Deriving loudness growth functions from categorical loudness scaling data

Marcin Wróblewski,[a] Daniel M. Rasetshwane, Stephen T. Neely, and Walt Jesteadt[b]
Boys Town National Research Hospital, Omaha, Nebraska 68131, USA

(Received 13 July 2017; revised 4 November 2017; accepted 28 November 2017; published online 19 December 2017)

The goal of this study was to reconcile the differences between measures of loudness obtained with continuous, unbounded scaling procedures, such as magnitude estimation and production, and those obtained using a limited number of discrete categories, such as categorical loudness scaling (CLS). The former procedures yield data with ratio properties, but some listeners find it difficult to generate numbers proportional to loudness and the numbers cannot be compared across listeners to explore individual differences. CLS, where listeners rate loudness on a verbal scale, is an easier task, but the numerical values or categorical units (CUs) assigned to the points on the scale are not proportional to loudness. Sufficient CLS data are now available to assign values in sones, a scale proportional to loudness, to the loudness categories. As a demonstration of this approach, data from Heeren, Hohmann, Appell, and Verhey [J. Acoust. Soc. Am. **133**, EL314–EL319 (2013)] were used to develop a $\mathrm{CU}_{\mathrm{sone}}$ metric, whose values were then substituted for the original CU values in reanalysis of a large set of CLS data obtained by Rasetshwane, Trevino, Gombert, Liebig-Trehearn, Kopun, Jesteadt, Neely, and Gorga [J. Acoust. Soc. Am. **137**, 1899–1913 (2015)]. The resulting data are well fitted by power functions and are in general agreement with previously published results obtained with magnitude estimation, magnitude production, and cross modality matching.

https://doi.org/10.1121/1.5017618

[JJL] Pages: 3660–3669

I. INTRODUCTION

Loudness has been defined as "that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from soft to loud" (ANSI, 2013, p. 58). Loudness is primarily a perceptual correlate of the physical strength of any given sound and has been quantified using a variety of measurement methods. In a recent review of measurement issues, Marks and Florentine (2011) note that scaling procedures can be described in terms of whether the scale is bounded or unbounded and whether it is discrete or continuous. Categorical scaling (CS) procedures are bounded and discrete, while procedures such as magnitude estimation are usually unbounded and continuous. Because scales developed with unbounded, continuous procedures are thought to have ratio properties (Stevens, 1972), making it possible to say that sound X is twice as loud as sound Y, we will refer to them as ratio-scale (RS) procedures. In the following paragraphs, we review the basics of CS and RS procedures, procedures for analysis of CS and RS data, and modification of a specific CS procedure to give it RS properties, making it possible to use the procedure to measure loudness in sones.

CS procedures have been widely used in the measurement of loudness, and several different procedures have been proposed for clinical use (Cox, 1989; Allen *et al.*, 1990; Kollmeier and Hohmann, 1995; Launer, 1995). Elberling (1999) noted the variable results obtained within and across similar CS procedures in listeners with normal hearing (NH). More recently, a version of the CS procedure described by Brand and Hohmann (2002) has been adopted as an ISO standard (ISO, 2006) and has been used in a number of studies, including studies of reliability (e.g., Al-Salim *et al.*, 2010; Heeren *et al.*, 2013; Rasetshwane *et al.*, 2015). Measures of reliability obtained with this procedure are generally better than the summary by Elberling (1999) would suggest. Only the categorical loudness scaling (CLS) procedure described in ISO (2006) will be discussed here, although the approach presented here could be used with all of the other CS procedures in the loudness scaling literature.

The ISO standard cites Brand and Hohmann (2002) as an example of a reference method for CLS. They used an 11-category scale, from not heard to extremely loud, which had been used in several of the earlier studies. Brand and Hohmann (2002) used initial trials to determine the lower and upper bounds of the range of stimulus intensities in a way that allowed them to include those trials in the final determination of the loudness function. The data were analyzed by assigning 11 numerical values, or categorical units (CUs), in steps of 5 from 0 to 50, to the 11 categories. Functions were then fitted to the data describing loudness in CU as a function of stimulus level. The data are not well fitted by a single straight line, and Brand and Hohmann (2002) proposed a two-line fit with a steeper function at higher levels and a smooth Bezier-fit transition between the two lines. Alternative functions and fitting procedures are described by Brand (2000), Oetting *et al.* (2014), and Trevino *et al.* (2016). None of them are straightforward.

CLS has been used recently to explore the effects of stimulus duration (Valente *et al.*, 2011), bandwidth

[a] Also at: Pacific University, Hillsboro, OR 97123, USA.
[b] Electronic mail: walt.jesteadt@boystown.org

(Hots *et al.*, 2014), and the combination of the two (Verhey and Kollmeier, 2002; Anweiler and Verhey, 2006), as well as to obtain equal-loudness contours (Heeren *et al.*, 2013; Rasetshwane *et al.*, 2015) and measures of binaural loudness summation (Oetting *et al.*, 2016). A study to assess the test-retest reliability of CLS data found that the slopes of straight-line functions fitted to the whole range of data were correlated across sessions with $r = 0.94$ at 2000 Hz and 0.80 at 1000 Hz (Al-Salim *et al.*, 2010) when listeners with NH and sensorineural hearing loss (SNHL) were included in the analysis. Rasetshwane *et al.* (2015) also addressed reliability, but reported different metrics than those used by Al-Salim *et al.* (2010). A reanalysis of their data using comparable measures found r values of 0.84 at 2000 Hz and 0.82 at 1000 Hz.

The early efforts to arrive at a numerical scale of loudness with ratio properties were based either directly or indirectly on procedures where subjects were asked to adjust stimulus levels to double or halve loudness (Richardson and Ross, 1930; Churcher, 1935). Fletcher and Munson (1933) assumed that loudness would be doubled when the stimuli were presented binaurally rather than monaurally or when two equally loud tones widely separated in frequency were presented simultaneously. Stevens (1936) combined the available data to propose a sone scale of loudness, where 1 sone was the loudness of a 1000 Hz tone presented binaurally in a free field at 40 dB sensation level (SL). Stevens (1955) later combined many sets of data, including the earliest magnitude-estimation data, to propose that loudness at 1000 Hz could be described as a power law with an exponent[1] of 0.3.

Magnitude estimation of loudness, a procedure in which subjects are presented with tones or narrowband noises differing in intensity and are asked to assign numbers proportional to loudness (Stevens, 1956), and magnitude production, where the subjects are given numbers and asked to adjust the intensity of a sound (Stevens, 1957), have been used in combination in later studies (e.g., Hellman and Zwislocki, 1961, 1963). These two procedures are sometimes combined with cross-modality matching (CMM; Stevens, 1959) in which subjects adjust the magnitude of another quantity, such as line length, to be proportional to loudness (Teghtsoonian and Teghtsoonian, 1983). Hellman (1999) summarizes results of all three of these procedures obtained in her classic studies of the growth of loudness in listeners with SNHL.

Hellman (1999) reported data obtained with magnitude-estimation and magnitude-production measures of loudness and CMM to line length for 83 listeners with NH and 128 with long-duration SNHL (see also Hellman and Meiselman, 1990; Hellman and Meiselman, 1993). The auditory stimuli consisted of 1-s tone bursts in the frequency range from 500 to 4000 Hz. Only one frequency was tested per listener and the data were not sorted by test frequency, presumably under the assumption that the slope of the loudness function is relatively uniform over that range. Test-retest reliability was assessed by retesting 36 listeners. The data were summarized in terms of exponents for the linear portion of the power function relating loudness to signal level. For listeners with SNHL, the range of the linear portion began at 4 dB above threshold and never extended beyond 30 dB above threshold. Hellman (1999) reported mean slopes for average SNHL of 0, 45, 55, 65, and 75 dB HL. The data show steeper functions with greater SNHL, as would be predicted from results obtained in loudness matching studies (e.g., Miskolczy-Fodor, 1960; Hellman and Zwislocki, 1964; Moore and Glasberg, 1996).

The data from CLS and RS procedures are not plotted in the same coordinates and cannot be interpreted in the same framework. CLS data are analyzed and plotted by assigning arbitrary numerical values from 0 to 50, referred to as categorical units or CUs, to the 11 categories. The data points are the average (typically median) levels assigned to a given category, so the variability in measurement occurs on the x axis rather than on the y axis. The best fitting lines, with slopes described in terms of dB/CU, are typically obtained by minimizing squared deviations on y. Oetting *et al.* (2014) have noted that it would be preferable to minimize squared deviations on x when using CLS for hearing aid fitting because the goal is to minimize error in predicting the gain required for a given loudness category. The resulting functions are typically summarized in terms of the low slope, the high slope, and the level associated with the breakpoint between the two functions. CLS functions are generally based on data obtained in 20–40 trials. A number of approaches have been described to reduce the impact of outliers and otherwise reduce the variability in the resulting parameter estimates. Oetting *et al.* (2014) proposed incorporating information regarding quiet thresholds and assuming a fixed slope for the upper portion of the function if there were too few trials to obtain a slope estimate. Rasetshwane *et al.* (2015) computed a median level for all presentations with the same CU rating, and then excluded trials more than 12 dB from the median before recomputing the median.

Like CLS, RS procedures typically use small numbers of trials. Because the magnitude estimates are assumed to have RS properties and variability is assumed to be proportional to overall loudness, the data are analyzed by computing geometric means across loudness estimates at each level, then fitting a linear function in log sones. Subjects sometimes have to be re-instructed to use numbers proportional to loudness rather than a more limited range, but outliers are rarely a problem. Conversion of RS data to a sone scale is straightforward. Since the loudness of a 40-phon[2] tone (presented binaurally in a free field) is 1 sone, the numbers assigned to tones presented at a series of levels can be divided by the number assigned to a tone presented at 40 phons obtained from the best fitting power-law function. This changes the intercept of the function, but not the slope.

The use of a linear function is now understood to be an approximation (Florentine and Epstein, 2006). The function at 1000 Hz becomes steeper below 40 dB sound pressure level (SPL; see Jesteadt and Leibold, 2011) and above 100 dB SPL (Viemeister and Bacon, 1988) in listeners with NH. Hellman and Meiselman (1990, 1993) showed functions for SNHL listeners that were steeper than normal at low levels, but curved down at high levels as the growth of loudness with level becomes more like that of NH listeners. An example is shown in Fig. 1.

Category rating tasks are easier to explain to subjects than RS tasks and they do not require subjects to work with numbers. The primary drawback of CLS is that the resulting data do not provide measures of loudness growth that can be related to loudness models or to loudness data obtained with other procedures. Differences in the results obtained with CLS and RS tasks are sometimes attributed to fundamental limitations in category rating procedures. Marks and Florentine (2011) provide an excellent summary. They note that successive categories are typically assigned successive (or equally spaced) integer numbers for purposes of analysis, as if the categories represented an interval scale. The use of categories is heavily influenced by the range and distribution of the stimuli to be rated (Parducci, 1965). Stevens and Galanter (1957) summarized many comparisons showing a curvilinear relation between category and ratio scales and noted that category scales were more sensitive to stimulus spacing (Stevens, 1958). More recent attempts to compare CLS and RS results in the same subjects have found that exposure to the CLS task greatly alters behavior in a RS task, but that the reverse is not true (Blum *et al.*, 2000; Jesteadt and Joshi, 2013).

Heeren *et al.* (2013) obtained CLS data from 31 listeners with NH for third-octave bands of noise at 9 center frequencies between 250 and 8000 Hz and fitted functions to the data at each frequency as described by Brand and Hohmann (2002). They then obtained the levels corresponding to given loudness values in CUs from the individual functions and used the median level for each CU value, across the 31 subjects, to obtain a median loudness function. Equal-loudness contours were constructed by plotting the median values as a function of frequency. Heeren *et al.* (2013) developed a transform from sones to CUs, as described below, to be used to convert the output of loudness models from sones to CUs for comparison to CLS data.

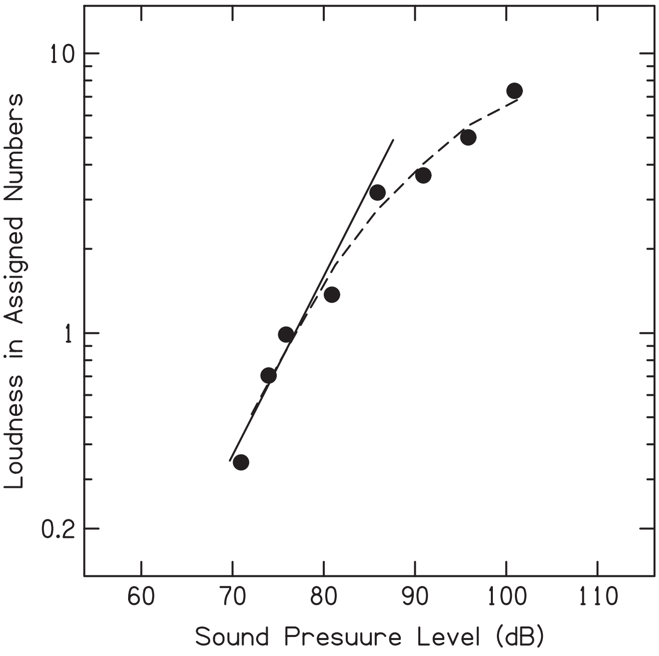

FIG. 1. Example replotted from Hellman and Meiselman (1990) showing curvature of the loudness function at high levels for listeners with cochlear impairment, and their estimation of a power-law exponent by fitting a straight line to the lower portion of the curve. The power-law exponent for the solid straight line is 0.64.

Rasetshwane *et al.* (2015) used similar procedures, consistent with the ISO standard, to obtain CLS data for pure tones at multiple frequencies in a large group of subjects (61 with NH and 87 with SNHL). In the present paper, we have used the conversion from sones to CUs developed by Heeren *et al.* (2013) to develop the opposite conversion, from CUs to sones, and have applied that conversion to the data obtained by Rasetshwane *et al.* (2015). Our reanalysis of the data suggests that substituting values in sones for the current arbitrary CUs would have many advantages for using CLS as a tool for the measurement of loudness.

II. METHODS

Loudness scaling data reported by Rasetshwane *et al.* (2015) have been used for all analyses conducted in the current study. Procedures used to obtain the data are reviewed here. Additional detail and rationale regarding stimuli, instrumentation, and data collection measurement procedures can be found in Rasetshwane *et al.* (2015).

A. Participants

The data were collected from 148 listeners. Sixty-one listeners with NH ranged in age from 11 to 53 yr with a mean age of 28.9 yr, standard deviation (SD) = 10.7, had thresholds equal to or better than 15 dB Hearing Level (HL) at audiometric frequencies. Eighty-seven listeners with SNHL ranged in age from 13 to 75 yr with a mean age of 55 yr, SD = 17.6, had mild to profound SNHL for at least one audiometric frequency. All participants had normal tympanometric middle ear compliance (>0.2 mmho), and had no air bone gaps greater than 10 dB. To assess test-retest reliability, 22 participants repeated the study.

B. Measurement and stimuli

Loudness data were obtained using an adaptive CLS procedure that determined the level of pure tones corresponding to different loudness categories (Brand and Hohmann, 2002; ISO, 2006). The pure-tone stimuli were 1000 ms long with rise/fall times of 20 ms. Data were collected monaurally and separately for each audiometric frequency. Octave and inter-octave audiometric frequencies between 250 and 8000 Hz were tested for listeners with NH. Listeners with SNHL judged loudness at all octave audiometric frequencies, and at inter-octave frequencies between two adjacent octave frequencies with threshold difference of 20 dB or more. Only octave frequencies were analyzed in the current study. Each listener completed 3 blocks of at least 11 stimulus presentations per frequency, following a practice run at 1250 Hz. The starting presentation level was 60 dB SPL for NH listeners, and varied for listeners with HL, depending on their degree of SNHL.

III. RESULTS

A. Conversion to sones

Heeren *et al.* (2013) provide a five-parameter formula for conversion from sones to CUs, replotted in Fig. 2. Their underlying assumptions were that the median levels in dB SPL for the 11 CLS categories at 1000 Hz could be interpreted as loudness levels in phons and that the phons could be converted to sones using a function describing the growth of loudness at 1000 Hz. There are many alternative functions available in the literature (reviewed by Jesteadt and Leibold, 2011 and Marozeau, 2011). The loudness function used by Heeren *et al.* (2013) from the ANSI (2007) standard, based on the Moore *et al.* (1997) loudness model, predicts that loudness is a power law of intensity with an exponent of 0.31, for levels from 40 to 100 dB SPL. The Heeren *et al.* (2013) formula is a 5-parameter polynomial that cannot be inverted to provide estimates of sones from CUs, but it is a simple matter to determine the sone values that correspond to CU values from 5 to 50 in steps of 5. A simple linear search, as implemented by the Goal Seek function in Excel, was used to obtain the values shown in Table I. No other values are necessary for the reanalysis of CLS data.

Note the absence of the CU of 0 (i.e., "cannot hear"). Sounds at or even just below threshold may have loudness (e.g., Moore *et al.*, 1997; Buus *et al.*, 1998), but sounds in this category are below threshold by an unknown amount. Similarly, in our analyses we chose to exclude CU of 50 (i.e., "too loud" or "extremely loud") and the corresponding $\mathrm{CU_{Sone}}$ value due to the relatively broad range of that loudness category.

After simply substituting the $\mathrm{CU_{Sone}}$ values in Table I for the original CU values, the CLS data at 1000 Hz were analyzed using two approaches. The first approach followed the procedures described in Rasetshwane *et al.* (2015) except for the final stage where they fitted CUs with two linear functions. Specifically, outliers were removed and median SPL for each CU was calculated. At the first stage of the analysis, for each CU the data that deviated more than 12 dB from the median SPL were removed, and the remaining scores were used to calculate new median SPL for a given CU. At the second stage of the analysis, any non-monotonic median SPL values (i.e., values smaller than +1 dB between adjacent increasing CU categories) were removed, to make the resulting CLS functions monotonic, as an increase in intensity should result in increase in loudness. The remaining median SPL values and corresponding loudness $\mathrm{CU_{sone}}$ values were then fitted with a power-law function, minimizing squared deviations in log $\mathrm{CU_{sone}}$ values.

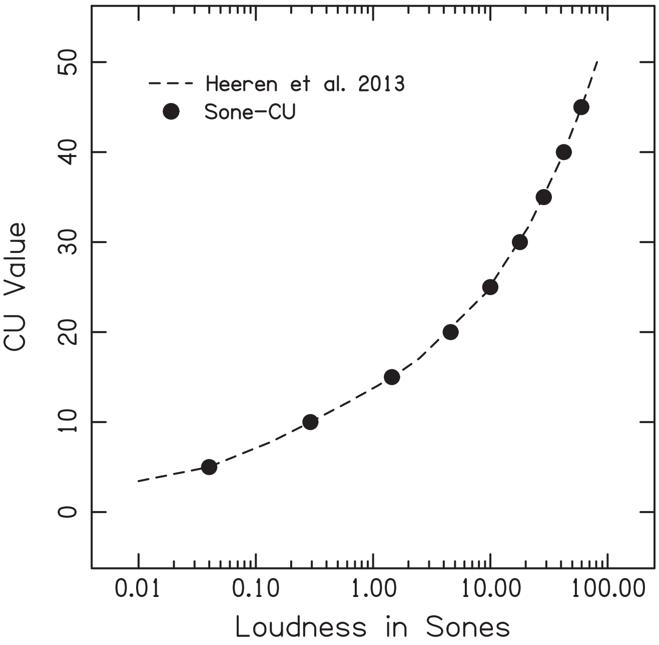

FIG. 2. CU to $\mathrm{CU_{sone}}$ transform. The five-parameter polynomial function from Heeren *et al.* (2013), plotted as a dashed line, is given by $\mathrm{CU} = 2.6253\log_{10}(\mathrm{sone}+0.0887)^3 + 0.7799\log_{10}(\mathrm{sone}+0.0887)^2 + 8.0856\log_{10}(\mathrm{sone}+0.0887) + 13.4493$. The filled symbols mark the CUs. Vertical lines from those points specify the corresponding $\mathrm{CU_{sone}}$ values (see Table I).

TABLE I. CLS category labels, CU values, and corresponding transformed $\mathrm{CU_{sone}}$ values.

Label	CU	$\mathrm{CU_{sone}}$
Too loud	50	81.06
Very loud	45	59.71
	40	42.35
Loud	35	28.54
	30	17.87
Medium	25	9.99
	20	4.58
Soft	15	1.45
	10	0.29
Very soft	5	0.04
Cannot hear	0	N/A

The second approach involved no data manipulation, as raw trial-by-trial data were used for all analyses. In this approach, individual trials were recorded as the level of the tone in dB SPL and the corresponding loudness judgment using $\mathrm{CU_{sone}}$ values. The data for all trials for a given frequency and subject were then fitted with a power-law function, minimizing squared deviations in log $\mathrm{CU_{sone}}$ values. This analysis was modeled on procedures used in fitting magnitude estimation data. Because plots of loudness in log sones vs level in dB SPL typically show curvature below 40 dB SPL at 1000 Hz, the power-law fits obtained with both approaches were tested for curvature by evaluating the pattern of deviations from the best fitting line and omitting the lowest CU category and refitting the data. Neither test showed evidence of curvature. The same tests were used to evaluate curvature at high levels in listeners with SNHL, as illustrated in Fig. 3. Again, neither test showed evidence of curvature.

These two approaches yielded comparable results at 1000 Hz, with a correlation between the two sets of slopes of $r = 0.96$. The data at frequencies other than 1000 Hz were therefore analyzed using only the second approach (viz., using raw trial-by-trial data). The slopes of loudness growth were obtained for all listeners and all octave frequencies tested by Rasetshwane *et al.* (2015), a total of 888 loudness functions.

B. Slope as function of hearing loss

Use of the ANSI (2007) standard, or any similar function, to convert from phons to sones ensures that the growth

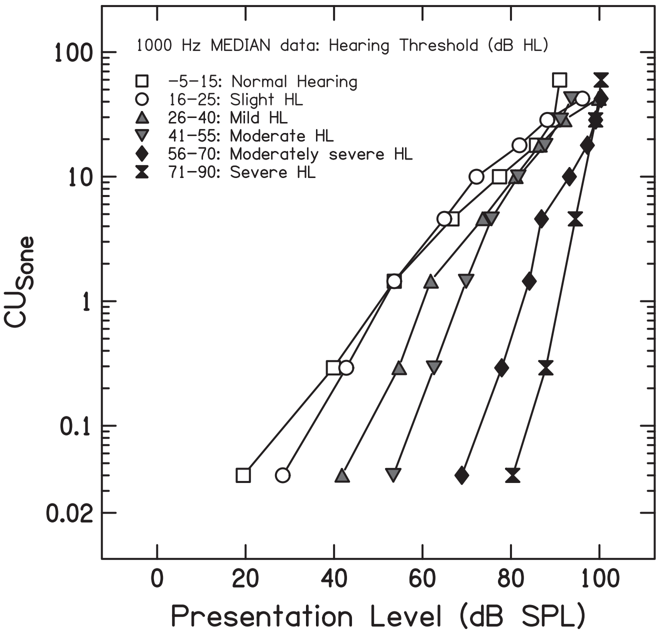

FIG. 3. Mean data for growth of loudness as a function of level for each of six HL groups. Each data point is the mean across all listeners in a given group of the final median SPL values associated with each CLS category.

of loudness at 1000 Hz in listeners with NH will be described by a power law with an exponent of ∼0.3, because that is assumed in the standard. The best test of the value of this modification of the CLS procedure is to apply it to conditions not constrained by the assumed loudness function and to compare the results to those obtained with RS methods. The data from 87 listeners with SNHL obtained by Rasetshwane *et al.* (2015) therefore provide an interesting test. Recruitment functions based on the first approach described above, using median SPL values for each CLS category, are shown in Fig. 3. The mean data show little of the curvature at high levels shown in Fig. 1, but some curvature is apparent in the lower and middle hearing loss categories.

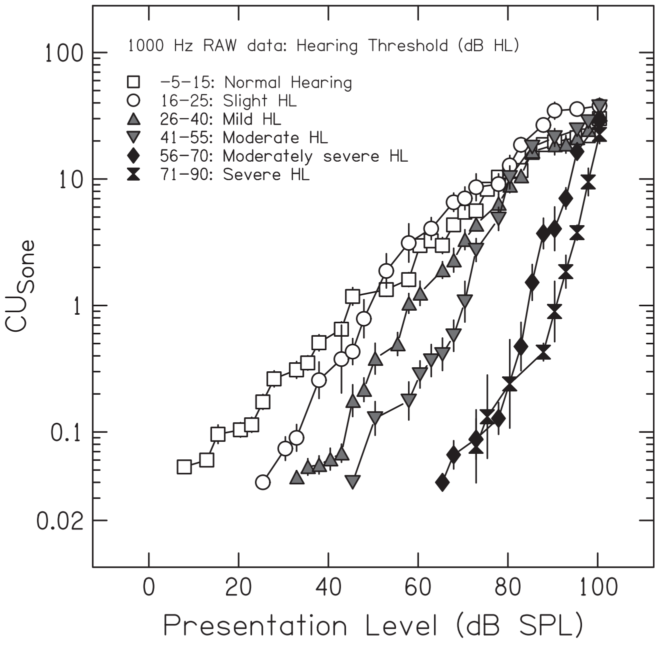

FIG. 4. Geometric means and standard errors for growth of loudness as a function of level for each of six SNHL groups based on the raw data. The figure shows 110 mean values. An additional 60 non-monotonic mean values (representing 17% of the trials) were not included in the figure. Most, but not all, removed data points were based on a low number of trials and were at either extreme of the loudness growth functions.

The transformation from CU to $\mathrm{CU}_{\mathrm{sone}}$ metric makes it possible to determine the geometric mean value of $\mathrm{CU}_{\mathrm{sone}}$ at every SPL represented in the raw data. Mean recruitment functions obtained that way are shown in Fig. 4.

The raw CLS data for 61 individual listeners with NH and 87 individual listeners with SNHL obtained by Rasetshwane *et al.* (2015) were analyzed using the second approach described above, by computing the geometric mean value in sones for each individual level in dB SPL presented to the listener (as shown for mean data in Fig. 4) and fitting a power function that minimized deviations on the *y* axis. Geometric mean exponents were then computed for all data points, including those omitted from Fig. 4, for each hearing loss group to allow a comparison to data for 128 listeners with NH and cochlear impairment summarized by Hellman (1999). The results are shown in Fig. 5. The data point at 90 dB HL represents results for one listener.

The power-law slopes obtained at 1000 Hz by reanalysis of CLS data grow less rapidly with hearing loss than the slopes reported by Hellman (1999), but part of the difference is no doubt due to the curvature shown in Fig. 1. The slopes reported by Hellman (1999) would have been shallower if

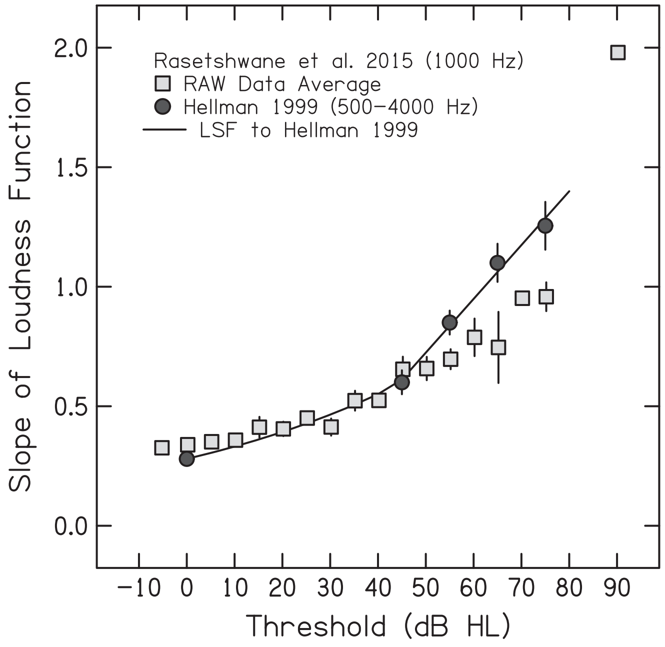

FIG. 5. Slopes of power-law fits to the transformed raw data obtained by Rasetshwane *et al.* (2015) at 1000 Hz compared to summary provided by Hellman (1999) in Fig. 8 of that paper. She showed data from Hellman and Meiselman (1990), filled circles with black line, for a group of listeners with NH and four groups with average SNHL. of 45, 55, 65, and 75 dB. The test frequency in her data varied from 500 to 4000 Hz. The Hellman slope values have been reduced by a factor of two to show them in power rather than pressure. The straight line plotted by Hellman has a slope of 0.0225 per dB of hearing loss. She drew a curved line from the point at 0 dB HL to the point at 45 dB HL, but lacked data over that range.

more points had been included in the power-law fit and, likewise, slopes obtained in the reanalysis of CLS data would have been steeper if some data points had been excluded. Other factors could be that the CLS scale is restricted at the high end or that listeners with SNHL do not interpret the verbal labels in the same way as listeners with NH. Etiology might also play a role. All of Hellman's listeners reported a history of noise exposure, whereas only six of Rasetshwane's listeners reported noise exposure and the etiology in most cases was unknown. Both sets of data show an orderly increase in recruitment as a function of increasing SNHL.

C. Results at other frequencies

The analysis of raw data at other frequencies yielded results comparable to those at 1000 Hz. Figure 6 shows slopes of power-law fits to the transformed raw data obtained by Rasetshwane *et al.* (2015) for octave frequencies between 250 and 8000 Hz juxtaposed with summary data for frequencies in the range of 500–4000 Hz from Hellman (1999). At thresholds between 40 and 80 dB HL, the slopes of power-law fits for 500 and 8000 Hz most closely approximate those from Hellman (1999). At 250 and 2000 Hz they become shallower. Frequencies of 1000 Hz, and particularly 4000 Hz, show the shallowest pattern of slope growth with increasing threshold for data from Rasetshwane *et al.* (2015) relative to Hellman (1999). As mentioned above, the discrepancy between the slopes we derived from Rasetshwane *et al.* (2015) and the data reported by Hellman (1999) stems from limited number of points included in the power-law fit in the latter study, as well as our decision to not exclude any data points obtained in the former study. At all frequencies, both sets of data have a systematic increase in recruitment as a function of increasing hearing threshold.

D. Reliability

Hellman (1999) reported a correlation with $r = 0.83$ between CMM slope values from the test and retest for 36 listeners with HL. Recent CLS studies show similar reliability, but the differences in fitting procedures and additional parameters in the CLS studies preclude a direct comparison. Data from 22 listeners who returned for a second testing session in the study by Rasetshwane *et al.* (2015) were used to assess the repeatability of loudness growth measures and, more specifically, the slopes of the loudness function derived from raw data obtained using the CLS procedure. Reliability for $\mathrm{CU_{sone}}$ data at audiometric octave frequencies between 250 and 8000 Hz were analyzed across two testing sessions. For direct comparison with findings from Hellman (1999), Pearson product moment correlation coefficients (r) were computed for test and retest slope data, as shown in Fig. 7.

In comparing the values in Fig. 7 to those reported by Hellman (1999), it should be noted that her CMM data were obtained in a relatively short session, comparable to the time required for CLS, but her subjects were also tested in

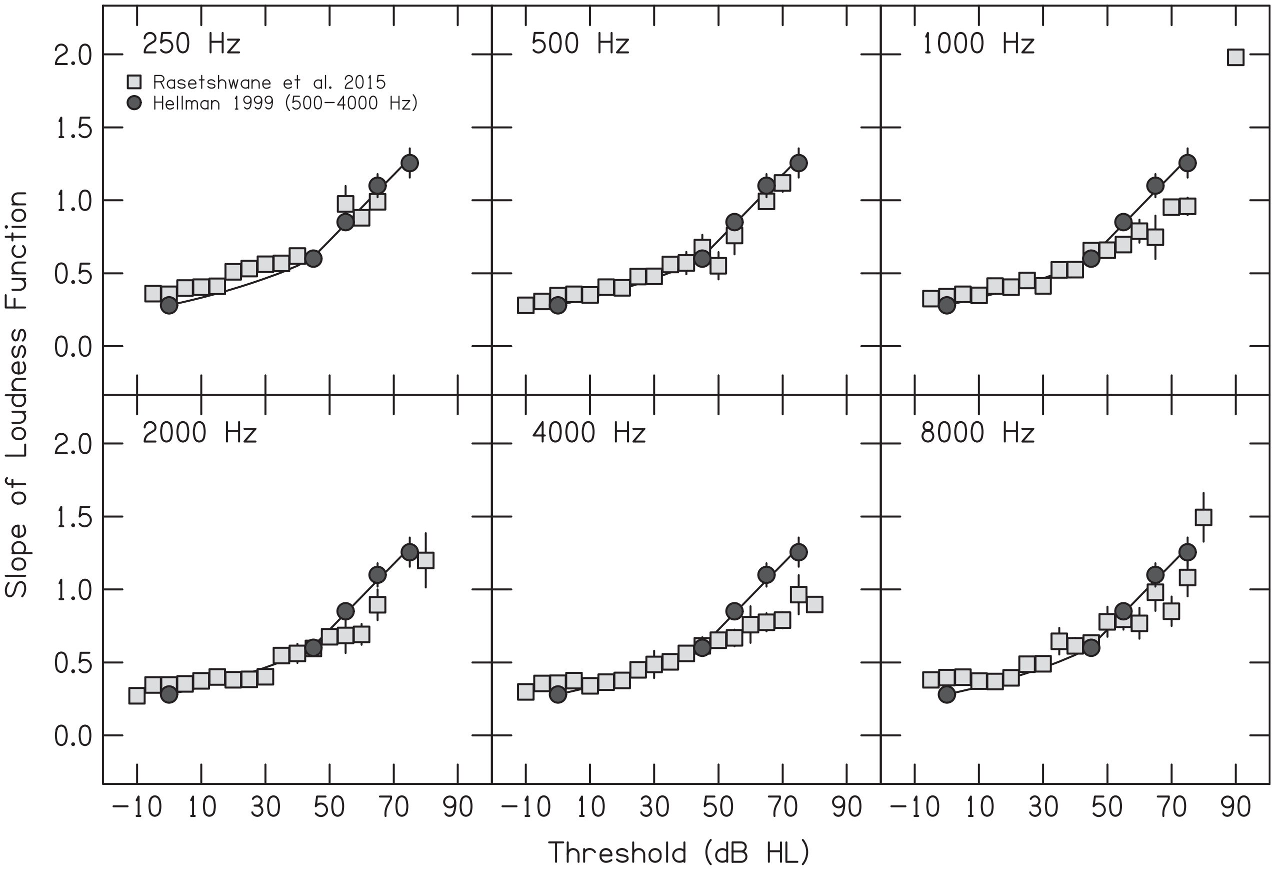

FIG. 6. Slopes of power-law fits to the transformed raw data obtained by Rasetshwane *et al.* (2015) at octave frequencies between 250 and 1000 Hz compared to summary provided by Hellman (1999) in Fig. 8 of that paper. The test frequency in her data varied from 500 to 4000 Hz.

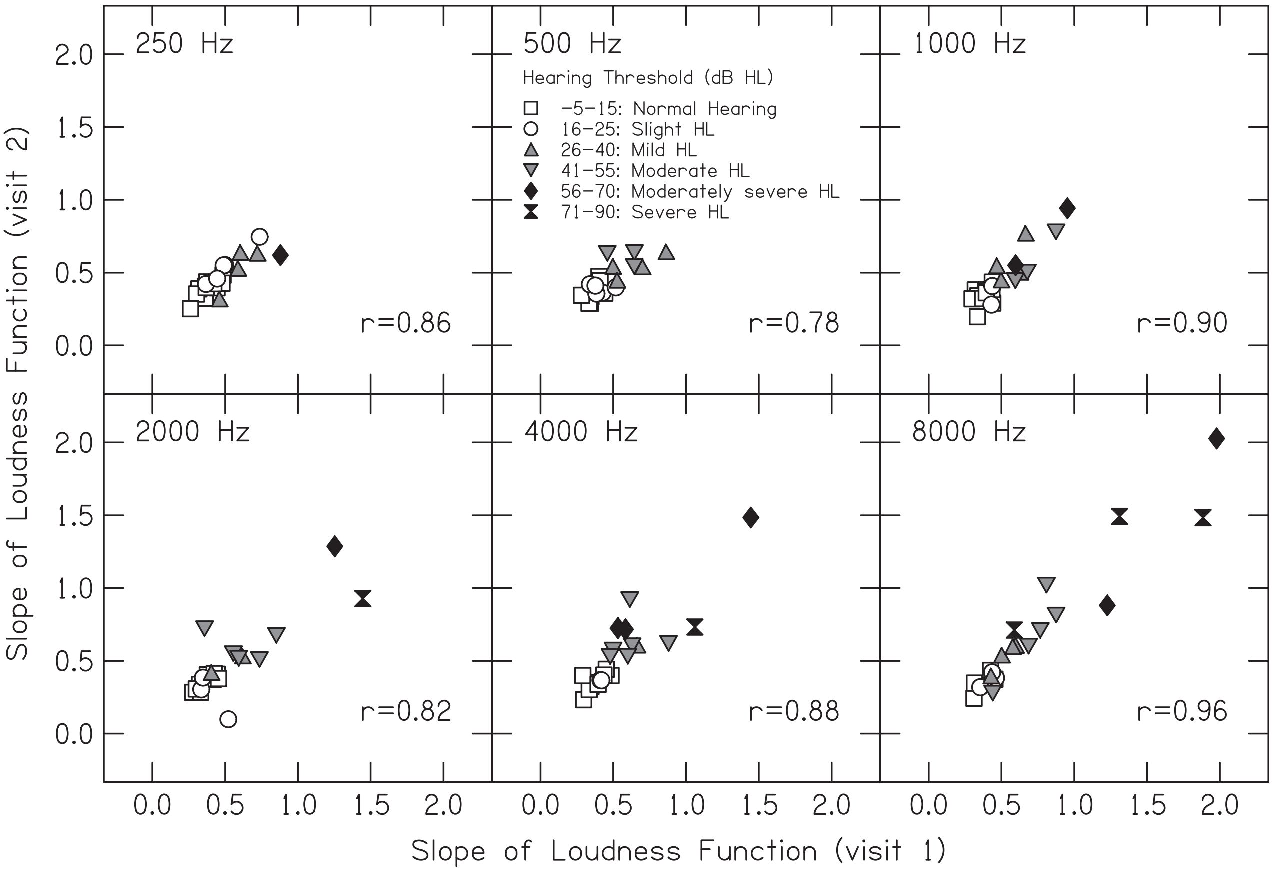

FIG. 7. Scatter plots showing the reliability of loudness function slopes derived by fitting power-law functions to the CU$_{sone}$ data for 22 listeners tested in 2 sessions by Rasetshwane *et al.* (2015). Hearing threshold (db HL) range is plotted as a parameter.

magnitude estimation and production in an effort to demonstrate the validity of the CMM measure. We are not aware of data demonstrating the reliability of CMM in isolation for listeners with SNHL.

IV. DISCUSSION

A. Limitations of the CU$_{sone}$ transform

Because Heeren *et al.* (2013) had fitted a function to their data, it was most straightforward to apply their function to the independent data set that we had available, but their data set was relatively small. It would be feasible to fit their function to larger sets of CLS data and arrive at a more valid set of CU$_{sone}$ values. A comparison of the Heeren *et al.* (2013) and Rasetshwane *et al.* (2015) data in Fig. 6 of Rasetshwane *et al.* (2015) shows that the average SPL values were lower in the latter study for lower CLS categories. This would result in shallower loudness growth functions when the Heeren *et al.* (2013) results were used to translate the Rasetshwane *et al.* (2015) data to CU$_{sone}$ values. Although the CU$_{sone}$ values in Table I could be considered preliminary, they are less arbitrary and more highly correlated with loudness than the original CU scale.

The assignment of CU$_{sone}$ values to CLS categories will not eliminate the problems associated with categorical procedures, such as the tendency to use categories equally often (Parducci, 1965), the sequential effects associated with categorical judgments (Trevino *et al.*, 2016), or the effects of stimulus spacing on categorical judgments (Stevens, 1958). Stevens (1956) found that magnitude estimates were not affected by the spacing of the stimuli. It should be noted, however, that a number of other context effects also occur in both CS and RS data (Arieh and Marks, 2011; Petzschner *et al.*, 2015).

Transformation of the CLS data from CU to CU$_{sone}$ made it possible to fit power-law functions to the data for individual listeners, but the resulting functions did not show the expected curvature below 40 dB SPL or the expected curvature in the loudness functions for listeners with SNHL as they approached the functions for listeners with NH at high presentation levels. The 11-point CLS scale, reduced to 9 points in the current analyses, may have limited the resolution of data.

Finally, it is clear that use of a CU$_{sone}$ scale will not provide new, independent information about the growth of loudness with intensity in listeners with NH. The scale assumes the current ANSI standard. There is still debate concerning the loudness function at 1000 Hz (Florentine and Epstein, 2006) and the normative value of the exponent for power-law fits above 1000 Hz. Many recent papers have reported exponents lower than 0.3 (Baird *et al.*, 1980; Viemeister and Bacon, 1988; Epstein and Florentine, 2005, 2006; Marozeau and Florentine, 2009). Marozeau *et al.* (2006) and Silva and Epstein (2010) describe corrections to low exponents obtained using CMM to string length under the assumption that the normative exponent is 0.3 and that low values lead to underestimates of the ratio of binaural to monaural loudness. The corrections are equivalent to relying on the current ANSI standard.

B. Advantages of the $\mathrm{CU}_{\mathrm{sone}}$ transform

Incorporation of $\mathrm{CU}_{\mathrm{sone}}$ values and power-law fits into adaptive CLS procedures such as ACALOS (Brand and Hohmann, 2002) would improve the distribution of stimulus levels over the dynamic range because the choice of levels would be based on a simpler, more accurate approximation to the loudness function. This would result in a more even use of the category labels. We have demonstrated that it is feasible to process the raw data without the need to determine the median level for each CLS category. Because there is no need for multiple occurrences of each category to arrive at a median, it is feasible to modify the procedure to increase the number of categories. It would be possible, for example, to instruct subjects to click on a position that was higher on a given category bar if they found the stimulus to be at the upper end of a given category and to treat position on the series of bars as a continuous scale. This would increase the number of effective categories, thereby reducing the differences between CLS and RS results (Stevens and Galanter, 1957; Marks and Florentine, 2011) although normative data would have to be obtained with any modified procedure.

While there are advantages for future data collection, the real benefits would occur in data analysis and in interpretation of the results. The substitution of $\mathrm{CU}_{\mathrm{sone}}$ values in place of the values discussed in the ISO (2006) standard would bring CLS data in line with RS data in the auditory research literature, allowing the CLS data to be summarized by a single number describing the power-law exponent or by two numbers specifying the slope and intercept of the best-fitting function. This would facilitate comparison of CLS data with loudness models. The correspondence between RS data and CLS data transformed to the $\mathrm{CU}_{\mathrm{sone}}$ metric is roughly equivalent to demonstrating that $\mathrm{CU}_{\mathrm{sone}}$ values meet the requirements of RS data and simplifies the analysis of CLS data. The step of obtaining medians is unnecessary and the data can be fitted by minimizing errors on the loudness axis rather than on input level. Because the assignment of numerical values to the CLS categories occurs after data collection, the conversion to $\mathrm{CU}_{\mathrm{sone}}$ values can be applied to existing data.

Because Rasetshwane *et al.* (2015) presented stimuli monaurally, we would expect the resulting power-law functions at 1000 Hz for listeners with NH to have values of 0.5 sone at 40 dB SPL. That can be tested by fitting a power-law function to the leftmost set of data in Fig. 4, summarizing results for 89 listeners with thresholds $\leq$15 dB HL at 1000 Hz. The function has an exponent of 0.31 and the estimated loudness at 40 dB SPL is 0.53 sone. This is another demonstration that the proposed conversion brings CLS data in line with the RS literature. This is an area, however, where current loudness standards are being reevaluated. Loudness models have been revised to reflect the fact that binaural loudness summation is not perfect under all conditions (Moore and Glasberg, 2007; Epstein and Florentine, 2012; ISO, 2017). The preliminary values in Table I could be revised to adjust for a difference in binaural loudness summation, resulting in a change in the intercept but not the slope of the loudness function.

The use of $\mathrm{CU}_{\mathrm{sone}}$ values facilitates development of loudness recruitment functions in listeners with SNHL. Loudness recruitment is often demonstrated by showing loudness matches over a range of levels for ears with NH and those with SNHL (e.g., Miskolczy-Fodor, 1960; Moore and Glasberg, 1997). This has also been done in studies using CS by plotting the levels associated with specific categories for ears with NH and those with SNHL (e.g., Allen *et al.*, 1990; Rasetshwane *et al.*, 2015). Hellman and Meiselman (1990) note that there are advantages to plotting the actual loudness functions of listeners with SNHL, as in Fig. 4, but that this has rarely been done outside of Hellman's own work. A number of studies have shown recruitment functions with level on the abscissa and CUs on the ordinate (e.g., Garnier *et al.*, 2000; Brand and Hohmann, 2002; Rasetshwane *et al.*, 2015; Kostek *et al.*, 2016; Oetting *et al.*, 2016), but those functions are all distorted by the transform between CUs and loudness. Use of $\mathrm{CU}_{\mathrm{sone}}$ values provides functions closer in form to those shown by Hellman and Meiselman (1990, 1993).[3] Although the recruitment functions for individual listeners lack curvature at high levels, use of more categories, or a continuous rating scale, which would be facilitated by use of $\mathrm{CU}_{\mathrm{sone}}$ values, might result in functions even closer to the example in Fig. 1.

It should be noted that RS methods such as magnitude estimation are well suited to studies of growth of loudness as a function of level, but do not lend themselves to comparisons across individuals, whereas CS methods using verbal labels such as "very loud" can be used to explore individual differences (Marks *et al.*, 1983). This is one reason that CS methods are widely used in clinical studies. Use of $\mathrm{CU}_{\mathrm{sone}}$ values preserves the meaning of the verbal labels while assigning values with ratio properties.

An examination of the $\mathrm{CU}_{\mathrm{sone}}$ values provides information about the values assigned to the verbal labels. A comparison of the values in Table I suggests that the label very loud is assigned to sounds that are twice as loud as those labeled "loud." The labels "medium" and loud are separated by a factor of about 3, while "soft" and medium are separated by a factor of 7, and "very soft" and soft are separated by a factor of 36. A comparison of the differences in sones rather than the ratios would show the opposite pattern, with about 1.4 sones separating soft and very soft, but 31 sones separating very loud and loud. The category labels are not evenly spaced in either sense. These relations should remain the same in the presence of loudness recruitment.

The unequal spacing of the category labels means that averaging CU values across multiple categories, as is sometimes done in reporting mean data, has the potential to distort the results. Studies of monaural and binaural loudness summation or loudness additivity would benefit from use of a measure where the raw data were proportional to loudness.

C. Generalization to other applications of CS

The basic idea underlying the proposed conversion is that sufficient information is available to assign meaningful values to loudness categories rather than relying on arbitrary values or on values determined by confusions between categories (i.e., Braida and Durlach, 1972). This is true only because extensive information on the growth of loudness as a function of intensity has been obtained with other measurement paradigms. This approach could be used, therefore, with other applications of CS to loudness (e.g., Allen *et al.*, 1990; Cox *et al.*, 1997), but not for applications of CS in general. There may be other continua, however, where extensive scaling data are available and category-rating procedures are desirable because of their ease of use. In those cases, it would be possible to use existing knowledge of the scale to assign meaningful numerical values to the categories.

Another approach to assigning numerical values to verbal category labels has been developed by Borg and colleagues (Marks *et al.*, 1983; Borg and Borg, 2001). They have used CMM to exertion and other techniques to assign perceptual intensities to verbal labels that were then used to develop "category-ratio" scales such that the verbal labels have ratio rather than linear properties. This way of assigning numerical values to verbal labels is more general and would represent an interesting converging operation if applied to the widely used CLS categories.

V. CONCLUSIONS

The CU-to-sone transform creates a set of $\mathrm{CU}_{\mathrm{sone}}$ values proportional to loudness, which helps overcome the arbitrariness of CUs and brings loudness functions obtained using the CLS procedure in line with those obtained with RS procedures such as magnitude estimation. The use of $\mathrm{CU}_{\mathrm{sone}}$ values would lead to greater flexibility in the design of data collection procedures, but could be used in the reanalysis of existing CLS data. For the data obtained by Rasetshwane *et al.* (2015), the resulting loudness functions are consistent with the recruitment pattern observed with increasing audiometric thresholds. The power-law exponents of the loudness functions closely approximate those obtained by Hellman (1999), who used more time-consuming methods, and show similar increase with increasing audiometric thresholds at all audiometric octave frequencies, although $\mathrm{CU}_{\mathrm{sone}}$ values for individual listeners do not show the expected curvature below 40 dB SPL and vary in the expected curvature in the loudness functions across frequencies for listeners with SNHL as they approach the functions for listeners with NH at high presentation levels. Procedures designed to "clean up" data prior to analysis are not necessary while using this quick and straightforward conversion, as raw data yields accurate and repeatable results. The high reliability of individual power-law exponents derived via $\mathrm{CU}_{\mathrm{sone}}$ values validates the use of this approach. Earlier work by Borg and colleagues provides an alternative way to arrive at category scales with ratio properties.

ACKNOWLEDGMENTS

This paper benefitted greatly from review by Josef Schlittenlacher and two anonymous reviewers. We are grateful to one anonymous reviewer for pointing out related work by Borg and colleagues. This work was supported by National Institutes of Health (NIH) Grant Nos. R01 DC011806, R03 DC013982, and P30 DC040662.

[1] The exponent corresponds to the slope of a line fitted to the loudness function when it is plotted as $10\log_{10}$(sones) vs dB SPL. We will use exponent and slope interchangeably.

[2] The phon scale provides a measure of loudness level such that a sound with a loudness level of x phons is equal in loudness to a 1000-Hz tone presented at x dB SPL.

[3] When growth of loudness in listeners with SNHL can be represented by power-law functions with higher than normal exponents, as in Fig. 4 or as in the exponents summarized in Fig. 5, the function describing loudness matches between ears with NH and SNHL will also be a straight line in dB coordinates with a slope given by the ratio of the power-law exponents. Thus, a power-law exponent of 0.6 rather than the exponent of 0.3 expected for an ear with NH corresponds to a matching function with a slope of 2, showing that a 2-dB increase in the normal-hearing ear will be required to match an increase of 1 dB in the ear with SNHL.

Allen, J. B., Hall, J., and Jeng, P. (**1990**). "Loudness growth in 1/2-octave bands (LGOB)—A procedure for the assessment of loudness," J. Acoust. Soc. Am. **88**, 745–753.

Al-Salim, S. C., Kopun, J. G., Neely, S. T., Jesteadt, W., Stiegemann, B., and Gorga, M. P. (**2010**). "Reliability of categorical loudness scaling and its relation to threshold," Ear Hear. **31**, 567–578.

ANSI (**2007**). S3.4-2007. *Procedure for the Computation of Loudness of Steady Sounds* (American National Standards Institute, New York).

ANSI (**2013**). S1.1-2013. *Acoustical Terminology* (American National Standards Institute, New York).

Anweiler, A. K., and Verhey, J. L. (**2006**). "Spectral loudness summation for short and long signals as a function of level," J. Acoust. Soc. Am. **119**, 2919–2928.

Arieh, Y., and Marks, L. E. (**2011**). "Measurement of loudness, Part II: Context effects," in *Loudness*, edited by M. Florentine, A. N. Popper, and R. R. Fay (Springer, New York), pp. 57–87.

Baird, J. C., Green, D. M., and Luce, R. D. (**1980**). "Variability and sequential effects in cross-modality matching of area and loudness," J. Exp. Psychol. Hum. Percept. Perform. **6**, 277–289.

Blum, R., Hohmann, V., and Kollmeier, B. (**2000**). "A comparison of categorical loudness scaling with the absolute magnitude estimation of loudness," Z. Audiol. **39**, 62–77.

Borg, G., and Borg, E. (**2001**). "A new generation of scaling methods: Level-anchored ratio scaling," Psychologica **28**, 15–45.

Braida, L. D., and Durlach, N. I. (**1972**). "Intensity perception. II. Resolution in one-interval paradigms," J. Acoust. Soc. Am. **51**, 483–502.

Brand, T. (**2000**). *Analysis and Optimization of Psychophysical Procedures in Audiology* (BIS Verlag, Oldenburg).

Brand, T., and Hohmann, V. (**2002**). "An adaptive procedure for categorical loudness scaling," J. Acoust. Soc. Am. **112**, 1597–1604.

Buus, S., Musch, H., and Florentine, M. (**1998**). "On loudness at threshold," J. Acoust. Soc. Am. **104**, 399–410.

Churcher, B. (**1935**). "A loudness scale for industrial noise measurements," J. Acoust. Soc. Am. **6**, 216–225.

Cox, R. M. (**1989**). "Comfortable loudness level: Stimulus effects, long-term reliability and predictability," J. Speech Hear. Res. **32**, 816–828.

Cox, R. M., Alexander, G. C., Taylor, I. M., and Gray, G. A. (**1997**). "The contour test of loudness perception," Ear Hear. **18**, 388–400.

Elberling, C. (**1999**). "Loudness scaling revisited," J. Am. Acad. Audiol. **10**, 248–260.

Epstein, M., and Florentine, M. (**2005**). "A test of the equal-loudness-ratio hypothesis using cross-modality matching functions," J. Acoust. Soc. Am. **118**, 907–913.

Epstein, M., and Florentine, M. (**2006**). "Loudness of brief tones measured by magnitude estimation and loudness matching," J. Acoust. Soc. Am. **119**, 1943–1945.

Epstein, M., and Florentine, M. (**2012**). "Binaural loudness summation for speech presented via earphones and loudspeaker with and without visual cues," J. Acoust. Soc. Am. **131**, 3981–3988.

Fletcher, H., and Munson, W. A. (**1933**). "Loudness, its definition, measurement and calculation," Bell. Syst. Tech. J. **12**, 377–430.

Florentine, M., and Epstein, M. (**2006**). "To honor Stevens and repeal his law (for the auditory system)," Proc. Fechner Day **22**, 37–42.

Garnier, S., Micheyl, C., Arthaud, P., and Collet, L. (**2000**). "Effect of frequency content on categorical loudness normalization," Scand. Audiol. **29**, 253–259.

Heeren, W., Hohmann, V., Appell, J. E., and Verhey, J. L. (**2013**). "Relation between loudness in categorical units and loudness in phons and sones," J. Acoust. Soc. Am. **133**, EL314–EL319.

Hellman, R. P. (**1999**). "Cross-modality matching: A tool for measuring loudness in sensorineural impairment," Ear Hear. **20**, 193–213.

Hellman, R. P., and Meiselman, C. H. (**1990**). "Loudness relations for individuals and groups in normal and impaired hearing," J. Acoust. Soc. Am. **88**, 2596–2606.

Hellman, R. P., and Meiselman, C. H. (**1993**). "Rate of loudness growth for pure tones in normal and impaired hearing," J. Acoust. Soc. Am. **93**, 966–975.

Hellman, R. P., and Zwislocki, J. (**1961**). "Some factors affecting the estimation of loudness," J. Acoust. Soc. Am. **33**, 687–694.

Hellman, R. P., and Zwislocki, J. (**1963**). "Monaural loudness function at 1000 cps and interaural summation," J. Acoust. Soc. Am. **35**, 856–865.

Hellman, R. P., and Zwislocki, J. J. (**1964**). "Loudness function of a 1000-cps tone in the presence of a masking noise," J. Acoust. Soc. Am. **36**, 1618–1627.

Hots, J., Rennies, J., and Verhey, J. (**2014**). "Loudness of subcritical sounds as a function of bandwidth, center frequency, and level," J. Acoust. Soc. Am. **135**, 1313–1320.

ISO (**2006**). 16832:2006, in *Acoustics-Loudness Scaling by Means of Categories* (International Organization for Standardization, Geneva, Switzerland).

ISO (**2017**). 532-2:2017, in *Acoustics—Methods for Calculating Loudness—Part 2: Moore-Glasberg Method* (International Organization for Standardization, Geneva, Switzerland).

Jesteadt, W., and Joshi, S. N. (**2013**). "Reliability of procedures used for scaling loudness," Proc. Mtgs. Acoust. **19**, 050023.

Jesteadt, W., and Leibold, L. J. (**2011**). "Loudness in the laboratory, Part I: Steady-state sounds," in *Loudness*, edited by M. Florentine, A. N. Popper, and R. R. Fay (Springer, New York), pp. 109–144.

Kollmeier, B., and Hohmann, V. (**1995**). "Loudness estimation and compensation employing a categorical scale," in *Advances in Hearing Research*, edited by G. Manley, G. Klump, C. Köppl, H. Fasti, and H. Oeckinghaus (World Scientific, Singapore), pp. 441–451.

Kostek, B., Odya, P., and Suchomski, P. (**2016**). "Loudness scaling test based on categorical perception," Arch. Acoust. **41**, 637–648.

Launer, S. (**1995**). "Loudness perception in listeners with sensorineural hearing impairment," Ph.D. thesis, Oldenburg University, Oldenburg, Germany.

Marks, L. E., Borg, G., and Ljunggren, G. (**1983**). "Individual differences in perceived exertion assessed by two new methods," Atten. Percept. Psycho. **34**, 280–288.

Marks, L. E., and Florentine, M. (**2011**). "Measurement of loudness, Part I: Methods, problems, and pitfalls," in *Loudness*, edited by M. Florentine, A. N. Popper, and R. R. Fay (Springer, New York), pp. 17–56.

Marozeau, J. (**2011**). "Models of loudness," in *Loudness*, edited by M. Florentine, A. N. Popper, and R. R. Fay (Springer, New York), pp. 261–284.

Marozeau, J., Epstein, M., Florentine, M., and Daley, B. (**2006**). "A test of the binaural equal-loudness-ratio hypothesis for tones," J. Acoust. Soc. Am. **120**, 3870–3877.

Marozeau, J., and Florentine, M. (**2009**). "Testing the binaural equal-loudness-ratio hypothesis with hearing-impaired listeners," J. Acoust. Soc. Am. **126**, 310–317.

Miskolczy-Fodor, F. (**1960**). "Relation between loudness and duration of tonal pulses. III. Response in cases of abnormal loudness function," J. Acoust. Soc. Am. **32**, 486–492.

Moore, B. C., and Glasberg, B. R. (**1996**). "A revision of Zwicker's loudness model," Acta Acust. Acust. **82**, 335–345.

Moore, B. C., and Glasberg, B. R. (**1997**). "A model of loudness perception applied to cochlear hearing loss," Aud. Neurosci. **3**, 289–311.

Moore, B. C. J., and Glasberg, B. R. (**2007**). "Modeling binaural loudness," J. Acoust. Soc. Am. **121**, 1604–1612.

Moore, B. C., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness, and partial loudness," J. Audio. Eng. Soc. **45**, 224–240.

Oetting, D., Brand, T., and Ewert, S. D. (**2014**). "Optimized loudness-function estimation for categorical loudness scaling data," Hear. Res. **316**, 16–27.

Oetting, D., Hohmann, V., Appell, J.-E., Kollmeier, B., and Ewert, S. D. (**2016**). "Spectral and binaural loudness summation for hearing-impaired listeners," Hear. Res. **335**, 179–192.

Parducci, A. (**1965**). "Category judgment: A range-frequency model," Psychol. Rev. **72**, 407–418.

Petzschner, F. H., Glasauer, S., and Stephan, K. E. (**2015**). "A Bayesian perspective on magnitude estimation," Trends Cogn. Sci. **19**, 285–293.

Rasetshwane, D. M., Trevino, A. C., Gombert, J. N., Liebig-Trehearn, L., Kopun, J. G., Jesteadt, W., Neely, S. T., and Gorga, M. P. (**2015**). "Categorical loudness scaling and equal-loudness contours in listeners with normal hearing and hearing loss," J. Acoust. Soc. Am. **137**, 1899–1913.

Richardson, L., and Ross, J. (**1930**). "Loudness and telephone current," J. Gen. Psychol. **3**, 288–306.

Silva, I., and Epstein, M. (**2010**). "Estimating loudness growth from tone-burst evoked responses," J. Acoust. Soc. Am. **127**, 3629–3642.

Stevens, J. C. (**1958**). "Stimulus spacing and the judgment of loudness," J. Exp. Psychol. **56**, 246–250.

Stevens, S. S. (**1936**). "A scale for the measurement of a psychological magnitude: Loudness," Psychol. Rev. **43**, 405–416.

Stevens, S. S. (**1955**). "The measurement of loudness," J. Acoust. Soc. Am. **27**, 815–829.

Stevens, S. S. (**1956**). "The direct estimation of sensory magnitudes: Loudness," Am. J. Psychol. **69**, 1–25.

Stevens, S. S. (**1957**). "On the psychophysical law," Psychol. Rev. **64**, 153–181.

Stevens, S. S. (**1959**). "Cross-modality validation of subjective scales for loudness, vibration, and electric shock," J. Exp. Psychol. **57**, 201–209.

Stevens, S. S. (**1972**). "Perceived level of noise by Mark VII and decibels (*E*)," J. Acoust. Soc. Am. **51**, 575–601.

Stevens, S. S., and Galanter, E. H. (**1957**). "Ratio scales and category scales for a dozen perceptual continua," J. Exp. Psychol. **54**, 377–411.

Teghtsoonian, M., and Teghtsoonian, R. (**1983**). "Consistency of individual exponents in cross-modal matching," Percept. Psychophys. **33**, 203–214.

Trevino, A. C., Jesteadt, W., and Neely, S. T. (**2016**). "Development of a multi-category psychometric function to model categorical loudness measurements," J. Acoust. Soc. Am. **140**, 2571–2583.

Valente, D. L., Joshi, S. N., and Jesteadt, W. (**2011**). "Temporal integration of loudness measured using categorical loudness scaling and matching procedures," J. Acoust. Soc. Am. **130**, EL32–EL37.

Verhey, J. L., and Kollmeier, B. (**2002**). "Spectral loudness summation as a function of duration," J. Acoust. Soc. Am. **111**, 1349–1358.

Viemeister, N. F., and Bacon, S. P. (**1988**). "Intensity discrimination, increment detection, and magnitude estimation for 1-kHz tones," J. Acoust. Soc. Am. **84**, 172–178.