Journal of Experimental Psychology: Applied
2002, Vol. 8, No. 4, 211–221

1076-898X/02/$5.00 DOI: 10.1037/1076-898X.8.4.211

Magnitude Estimation of Conceptual Data Dimensions for Use in Sonification

Bruce N. Walker
Rice University

Sonifications must match listener expectancies about representing data with sound. Three experiments showed the utility of magnitude estimation for this. In Experiment 1, 67 undergraduates judged the sizes of visual stimuli and the temperature, pressure, velocity, size, or dollars they represented. Similarly, in Experiment 2, 132 listeners judged the pitch or tempo of sounds and the data they represented. In both experiments, polarity and scaling preference depended on the conceptual data dimension. In Experiment 3, 60 listeners matched auditory graphs to data created with the results of Experiment 2, providing initial validation of scaling slopes. Magnitude estimation is proposed as a design tool in the development of data sonifications, with the level of polarity preference agreement predicting mapping effectiveness.

In virtually every science classroom and laboratory, researchers gather, analyze, and attempt to determine patterns in data. In many cases, the data sets are not only huge but also multidimensional and rapidly changing. Therefore, researchers must use all of the resources available, both technical and perceptual, to display and interpret their scientific results. However, most data exploration tools are exclusively visual in nature. These tools fail to exploit the excellent pattern recognition capabilities of the human auditory system and exclude students and researchers with visual disabilities.

Sonification is the use of nonspeech audio to convey information such as that used in the interpretation of scientific results. Specifically, *data sonification* is "the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation" (Kramer et al., 1999, p. 3). That is, scientific data, of any sort, are used to change the parameters of a synthesized tone. The many real and potential benefits of sonification have been detailed elsewhere (e.g., Kramer et al., 1999; Walker, 2000). However, almost no research has been done to determine how to create sonifications for maximum effectiveness, and there is little theory and virtually no experimental evidence to guide sonification researchers and designers (although research by Barrass, 1998, and Walker, 2000, have been a strong start). Designers have generally used whatever "sounded good" or "made sense" to them.

Critical Questions in Sonification

There are three initial questions that must be addressed by sonification researchers. The first question is: Which sound parameter is best for representing some data, such as temperature? Sonification theory depends on at least some agreement among users about what sound attribute most effectively represents a given data dimension. A follow-up question is whether there are gradations of "goodness." There may be some mappings that are considered excellent, or very obvious, some that are considered to be acceptable, and some mappings that are agreed to be poor mappings.

The next question is: What are the best polarities for those mappings? Listeners might agree that pitch should increase in order to represent increasing temperature (defined here as a *positive mapping polarity*), whereas pitch should decrease in order to represent increasing size (a *negative mapping polarity*).

Once a designer decides which sound dimension to use to represent the data, the third question is: How much change is required in, say, the pitch of a sound in order to convey a given change in, for example, temperature? This psychophysical scaling function is critical if sonifications are to be used to make accurate comparisons and absolute judgments.

Mappings and Metaphors: Walker and Kramer (1996)

Perhaps the first study intended specifically to address the issue of data-to-display mapping choices in sonifications (Walker & Kramer, 1996; see also, Barrass, 1998) showed that it matters which auditory dimension is used to display a given data dimension. Walker and Kramer sonified a fictitious factory where undergraduates monitored the data dimensions of temperature, pressure, size, and rate, and then made speeded responses based on changes in the sounds. The four data dimensions were represented by the auditory dimensions of loudness, pitch, tempo, or onset time (i.e., attack time) in mapping arrangements that differed for each experimental group. The researchers chose one mapping ensemble that seemed likely to be the best or most intuitive, a second that seemed simply okay, a third that seemed as if it would actually be bad or counterintuitive, and a fourth arrangement that was essentially random. Surprisingly, the mapping ensemble that resulted in the best performance was not the "intuitive" ensemble, but rather the "bad ensemble." Even the "random" ensemble outperformed the supposed best ensemble. It is clear from these results that

This article is based on the author's dissertation at Rice University. The research was supported by National Science Foundation Grant IIS-9906818. Portions of this research were presented at ICAD 2000, the International Conference on Auditory Display, Atlanta, Georgia, April 2000.

Correspondence concerning this article should be addressed to Bruce N. Walker, who is now at the School of Psychology, Georgia Institute of Technology, Atlanta, Georgia, 30332-0170. E-mail: bruce.walker@psych.gatech.edu

designers who must rely on their own intuition, in the absence of any guiding theory, may not be making the best sonification choices.

Walker and Kramer (1996) also looked at whether there was a particular display dimension that best represented a given data type. Both pitch and loudness were acceptable for representing all of the data types studied, whereas tempo was only a mediocre display dimension overall. In general, onset time was a fairly poor dimension, but it was actually the best choice of display dimensions for representing size data, even in the face of its poor performance with the other data types. Further, pitch was not as effective as loudness for representing temperature, despite the common experience that hot things (e.g., a tea kettle) tend to make higher pitched sounds as they become hotter. Finally, tempo, which might seem naturally suited to represent "rate" information, was only moderately successful in that role. Thus, it is clear that the specific data-to-display mapping has a large impact on performance in a rapid-response sonification-monitoring task. Although Walker and Kramer (1996) studied mapping choices, they did not specifically study the issues of polarity or scaling.

Psychophysics and Magnitude Estimation

The psychophysical method of magnitude estimation has become a standard approach to scaling the relationship between an acoustic variable and its perceptual correlate (Hellier, Edworthy, & Dennis, 1995; Stevens, 1975), and it invariably results in a power function. It has become clear that magnitude estimation can determine the scaling function between acoustic parameters and all sorts of other variables. Stevens (1975) claimed that any two dimensions can be compared and therefore scaled with a power function. Researchers testing this assertion have created scaling functions for some "nonsensory" dimensions, such as matching the loudness of a tone to the level of racism attributed to certain acts, the pronounceability of trigrams, and the desirability of certain professions (Dawson & Brinker, 1971). These can be considered as conceptual, rather than as perceptual data dimensions, because they are not within any particular sensory modality. Hellier, Edworthy, and Dennis (1993, 1995; see also Edworthy, Hellier, & Hards, 1995; Edworthy, Loxley, & Dennis, 1991) have used magnitude estimation to scale the relationship between acoustic parameters and perceived urgency, further demonstrating that conceptual as well as perceptual variables can be represented with sounds in a measured way. As an example, Edworthy et al. (1995) examined mapping, polarity, and scaling (they did not use all those terms) for the relationship between pitch, speed, harmonicity, and rhythm, and 42 conceptual adjectives such as dangerous, jerky, safe, and heavy. They found that some sound dimensions (pitch, speed) were generally more effective and also that some adjectives (e.g., dangerous, urgent) were more salient than others. Further, changes in some display dimensions (e.g., increases in pitch) consistently correlated with perceived changes in a data dimension (e.g., increases in urgency). The authors highlighted practical applications for those results in the design of trend-monitoring (i.e., prewarning) cockpit sounds.

These results suggest that magnitude estimation could be used to determine the scaling function between acoustic parameters and other, more widely used "data-type" conceptual dimensions such as temperature or pressure. Similar to the warning guidance that has emerged from the work of Hellier, Edworthy, and their colleagues (e.g., Edworthy et al., 1995), magnitude estimation may help designers with decisions about preferred mappings, polarities, and scalings, which are necessary in developing effective data sonifications.

Experiment 1: Magnitude Estimation With Lines and Circles

The investigation began by using magnitude estimation with simple visual stimuli to examine conceptual data dimensions in a context where somewhat similar stimuli had been used in the past. The perceptual dimension of size (e.g., line length) was included as a calibration of the experimental procedure, as there is some literature on the magnitude estimation of line length (e.g., Stevens, 1975; Stevens & Guirao, 1963; see also M. A. Teghtsoonian, 1965). A study of the literature determined that experiments involving estimations of the lengths of lines have shown exponent values to be systematically just less than 1.0. For example, Teghtsoonian and Teghtsoonian obtained values of 0.93, 0.98, and 0.98 for apparent length of lines (M. Teghtsoonian & Teghtsoonian, 1971, 1983, Experiment 1; 1983, Experiment 2, respectively). These results say nothing about what to expect for the matching of line length to the conceptual data dimensions of temperature, pressure, and velocity that were included in the present study.

In contrast to the near-linear relationship typically obtained for estimations of line lengths, M. A. Teghtsoonian (1965, p. 392) reported that for two-dimensional stimuli "judged size increases somewhat more slowly than the stimulus." Ekman (1958) found an exponent of 0.86 for circles. M. A. Teghtsoonian (1965), in a more thorough examination of the factors affecting perceived size, obtained an exponent of 0.76 with circles. M. Teghtsoonian and Teghtsoonian (1971) later found an exponent of 0.69 using outline circles. Stevens and Guirao (1963) used solid squares as stimuli and obtained an exponent of 0.70 for perceptual estimations. Considering these results, two-dimensional visual stimuli (solid circles) were also included in the present experiment.

There has been very little work relating two-dimensional shapes to any sort of conceptual data dimensions. However, Williams (1956, Experiment 4) conducted a study of map symbols to determine what size symbols should represent armies of different sizes. Replotting the data provided by Williams for circles yields a regression slope of $m = 0.73$, which falls in the middle of the range of results found for simple size estimations for solid circles. Unfortunately, there is no way to know what, if any, particular data dimension Williams' participants had in mind, because they provided responses about circles that would have 2, 3, 5, or 10 times the "value" of the first circle. On the whole, it is not clear whether one should expect any difference between estimations of size and estimations of conceptual value at all, let alone differences between different conceptual data types.

Method

Participants

Complete details of all experiments are provided in Walker (2000). In all experiments reported here, undergraduate students from Rice University participated for course credit. All reported normal or corrected-to-normal vision and hearing, signed informed consent forms, and provided demographic details about age, gender, handedness, and number of years of

musical training. A total of 67 participants (22 men, 45 women; mean age = 19.3 years) completed Experiment 1.

Apparatus and Stimuli

Visual stimuli appeared on a 17-in. (43-cm) computer display set to a resolution of 1,024 × 768 pixels, typically viewed from a distance of 24 in. (61 cm). The line stimuli in this experiment were nine black bars, five pixels wide × 10, 20, 40, 60, 80, 100, 400, 500, and 600 pixels long and oriented either horizontally or vertically. The circle stimuli were nine black solid circles, with diameters of 10, 30, 50, 70, 100, 300, 400, 500, and 600 pixels. The stimuli were displayed one at a time in the center of the screen with a white background.

Design

The experimental design included data dimension (temperature, pressure, velocity, and size) and display dimension (horizontal lines, vertical lines, and solid circles). Each data dimension (e.g., temperature) was paired with one of the display dimensions (e.g., horizontal lines) for an entire block of trials (e.g., named temperature:horizontal lines). Each participant completed two blocks of trials, with the constraint that each participant view two different display dimensions and two different data dimensions. An example experiment might include a size:circles block followed by a pressure:horizontal lines block. The number of participants in each block type is included in Table 1.

Trial Structure and Task

Before each block of trials, the experimenter read aloud instructions such as the following, while the participants followed along on the computer screen:

> You will see a series of lines on the screen in random order. Your task is to indicate what temperature they would represent, by assigning numbers to them. For the first line, assign it any number of your choosing that represents a temperature. Then, for each of the remaining lines, estimate its "temperature," relative to your subjective impression. For example, if the second line seems to represent a temperature that is 10 times as hot as the first, then assign it a number that is 10 times bigger than the first number. If the line seems to represent a temperature that is one-fifth as hot, assign it a number that is one-fifth as large as the first number, and so on. You can use any range of numbers, fractions, or decimals that seem appropriate, so long as they are greater than zero.

On each trial, the participant saw one stimulus from the set being used for that block (e.g., horizontal lines) and entered a number for the subjective value (e.g., the temperature) of that stimulus. In a block of 18 trials, each of the nine stimuli was randomly presented twice, with the constraint that the largest or smallest stimulus in that set could not occur first (see R. Teghtsoonian & Teghtsoonian, 1978). Following a brief rest, the participant began the second block with new instructions that introduced different data and display dimensions.

Results

The responses from all of the participants for each trial block type (i.e., for each combination of data and display dimensions) were combined and then sorted by stimulus number. Some participants used small numbers in their range of responses (e.g., 0.01–10), whereas others used somewhat larger numbers (e.g., 20–3,000). Therefore, the geometric mean was calculated for all responses for each stimulus across participants in a given block. These mean estimation values were plotted against the actual stimulus values in log–log coordinates and fitted with a power function of the form $y = bx^m$. The slope of the fit line, m, indicates how much the perceived, or estimated value changes as the actual stimulus parameter changes. This simple and direct analysis is all that is required to compute the slope, m. It should be noted that in studies in which responses to a specific stimulus rather than just the slope of the function is of interest, it is important to account for different moduli and ranges used by different participants. Stevens presented a normalizing procedure to account for these factors (Lane, Catania, & Stevens, 1961; see also Engen, 1971). However, as Lane et al. pointed out (p. 163), this normalization has no effect on the slopes of the regression lines, so it was unnecessary for the purposes of the present experiments. That fact simplifies the magnitude estimation procedure, making it even more straightforward for designers to apply.

Slopes and Polarities

All of the slopes obtained in Experiment 1 are summarized in Table 1. Size estimations are shown in the rightmost column. In the case of length estimation versus actual line length, the regression slopes were both just less than 1.0 (horizontal: $m = 0.96$, $SE_m = 0.02$, $r^2 = .995$; vertical: $m = 0.95$, $SE_m = 0.03$, $r^2 = .992$). The estimated size of circles versus their actual areas (in square pixels) was also well-fitted by a power function, in this case with slope $m = 0.59$ ($SE_m = 0.01$, $r^2 = .997$).

In contrast, the responses for the conceptual data dimensions mapped to the straight-line stimuli resulted in slopes in the range of 0.47–1.0, depending on the conceptual data dimension involved.

Table 1
Summary of Results From Experiment 1

	Slope of regression line, inside 95% confidence interval			
Display dimension	Temperature	Pressure	Velocity	Size (of stimulus)
Horizontal lines	0.84 ≤ 0.86 ≤ 0.88 (n = 10)	0.67 ≤ 0.74 ≤ 0.81 (n = 12)	0.45 ≤ 0.47 ≤ 0.49 (n = 10)	0.92 ≤ 0.96 ≤ 1.00 (n = 12)
Vertical lines	0.83 ≤ 0.89 ≤ 0.95 (n = 12)	0.49 ≤ 0.53 ≤ 0.57 (n = 10)	0.96 ≤ 1.00 ≤ 1.04 (n = 10)	0.88 ≤ 0.95 ≤ 1.02 (n = 10)
Circles				
Positive polarity	—	0.51 ≤ 0.56 ≤ 0.61 (n = 9)	0.51 ≤ 0.56 ≤ 0.61 (n = 7)	0.57 ≤ 0.59 ≤ 0.61 (n = 12)
Negative polarity	—	−0.48 ≤ −0.95 ≤ −1.42 (n = 3)	−0.45 ≤ −0.62 ≤ −0.79 (n = 3)	—

The analysis of responses to the circle stimuli representing conceptual data yielded a more complex picture (see Table 1). When estimating conceptual dimensions, some participants responded to the circles with polarities different from the majority. For example, 9 viewers indicated that increasing size meant increasing pressure, whereas 3 viewers responded that increasing size meant decreasing pressure. For that reason, the data that reflected different polarities within a mapping were analyzed separately. Table 1 includes the number of participants whose responses figured into the calculation of each slope.

Discussion

Simple perceptual estimates of the size of the stimuli were very near to the results with similar stimuli reported in the literature, which validates the experimental method used here to implement the magnitude estimation paradigm. When participants made conceptual magnitude estimations, the slopes were nearly all different from the slopes obtained for the perceptual dimension. Further, some of the data-to-display pairings yielded nonunanimous polarities with circle stimuli. Taken together, these results demonstrate that the nature of the data being represented has a significant effect on the value estimations made for visual stimuli and that magnitude estimation provides both polarity and slope details.

Participants may use unique (and even erroneous) mental models to guide their conceptual estimations. One participant reported thinking of the circle as a two-dimensional balloon and figured that there would be lower pressure if the balloon got larger, hence the negative-polarity mapping. Other participants indicated that they had used whatever just "felt right" for each combination of data and display dimensions. In an applied data-display setting, this variability of strategies and mental models may lead to responses that are not what any one display designer might expect.

Experiment 2: Magnitude Estimation With Sound Stimuli

Experiment 1 showed that interesting differences can be found between the polarity preferences and scaling functions for different conceptual data dimensions. Experiment 2 sought to extend this finding to the auditory dimensions frequency and tempo, which are commonly used in sonifications but which have not been systematically studied when representing conceptual data. In addition to the data dimensions used in Experiment 1 (temperature, pressure, velocity, and size), the concept of "number of dollars" was included. This is another broadly used data dimension, of interest in many fields such as economics, where both students and researchers are becoming more interested in using sonification to discover trends in their data.

Magnitude Estimation of Pitch

The perceptual dimensions of pitch and perceived tempo served in this experiment as calibrations of their acoustic correlates. Stevens and his colleagues (Stevens & Volkmann, 1940; Stevens, Volkmann, & Newman, 1937) used the methods of fractionation and equisection to develop the Mel scale, relating perceived pitch to frequency. If the Mel scale (Stevens, 1975, Figure 61) is re-drawn with log–log axes, in keeping with the magnitude estimation paradigm, and if the central region of the frequency spectrum (e.g., 100–3200 Hz, used by many sonifications) is selected, the perceived change in frequency versus the actual change in frequency is fitted by a power function with a slope of 0.73. Studies that have used more modern magnitude estimation procedures and less musical tones have tended to produce a slope just slightly steeper than the Mel scale (Beck & Shaw, 1961, 1962, 1963). Thus, if a free-modulus magnitude estimation procedure is used with frequencies between 100 and 3200 Hz, the expected relationship between perceived pitch and frequency should produce near-traditional psychophysical scaling plots, with slopes in the range of 0.73–0.80.

Magnitude Estimations for Tempo

A review of the literature has not uncovered any magnitude estimation experiments involving tempo per se. Eisler (1976) has, however, compiled a list of 111 studies that have attempted to obtain scaling estimates for duration. Note that the perception of nonsyncopated tempo is highly related to the perception of the duration of the elements. Across all of the studies, Eisler (1976, p. 1157) concluded that "time perception is not veridical; though the collected exponents straddle unity, most of them are smaller than 1 . . . [A] value of .9 seems to come closest to the exponent of subjective duration." Hence, we may assume that the estimation slopes for perception of tempo should also be slightly less than 1.0.

Of course, neither the Mel scale nor Eisler's (1976) results provide any indication of what scaling might be obtained with conceptual data estimations for pitch or tempo. In a related line of research, though, Hellier, Edworthy, and colleagues have included "speed" in several of their studies of urgency and warning-related concepts (e.g., Edworthy et al., 1995; Hellier et al., 1993). Their results support the prediction that tempo (via speed) can be used to represent concepts in a systematic and scaled manner.

Method

Participants

A total of 132 students participated (40 men and 92 women; mean age = 19.5 years). (From the same participant pool as in Experiment 1.)

Stimuli

Each auditory stimulus was composed of a one-beat long pure sine wave tone, followed by a half-beat of silence. These sound and silence elements were looped to create a continuous on–off pattern. Note that the length of a beat when measured in milliseconds depends on the tempo at which the sound is repeated. At 60 beats per minute (bpm), one beat lasts 1 s. There were two sets of stimuli: The 10 sounds in the frequency set were synthesized with tone frequencies of 100, 200, 300, 400, 800, 1000, 1400, 1800, 2400, and 3200 Hz, but were all played at a tempo of 60 bpm. The stimuli in the frequency set were normalized for perceived loudness, matching the 1000-Hz tone at 60 dBA SPL. The 10 sounds in the tempo set were all synthesized with a frequency of 1000 Hz but were repeated at tempos of 45, 60, 105, 150, 210, 270, 420, 500, 550, and 600 bpm. All of the sounds in the tempo set were presented at 60 dBA SPL. All sounds were presented via Sony MDR-V200 headphones. Finally, for the one block of participants who saw visual line stimuli rather than hearing auditory stimuli, the horizontal lines were identical to those used in Experiment 1.

Design

In this experiment, pitch and perceived tempo were the perceptual dimensions, whereas temperature, pressure, velocity, size, and number of

dollars were the conceptual data dimensions. The display dimensions were frequency and tempo. Each data dimension (e.g., temperature) was paired with one of the display dimensions (e.g., frequency) for an entire block of trials. The perceptual dimension of pitch was only paired with frequency, and the perceptual dimension of perceived tempo was only paired with tempo, because those data dimensions were calibrations of their respective auditory dimensions. As before, participants judged each stimulus twice within a block and completed two blocks of trials separated by a brief break. On each trial, the participant heard one stimulus from the set being used for that block (e.g., sounds from the frequency set).

As an initial test of the stability of the results from Experiment 1, 16 additional participants completed the magnitude estimation procedure between horizontal line length and perceived length. This was an exact replication of one of the visual trial block types from Experiment 1. Those data were gathered at the beginning of a separate and unrelated experiment, and those participants did not complete any blocks of trials using auditory stimuli.

Results

Horizontal Line Stimuli: A Replication

All results for this experiment are presented in Table 2. The block of trials comparing perceived and actual length of horizontal lines resulted in a nearly exact replication of Experiment 1. In the present experiment, the regression slope $m = 0.98$ ($SE_m = 0.02$), and $r^2 = .996$.

Auditory Stimuli: Individual Analyses of Polarity

The results with circle stimuli in Experiment 1 demonstrated individual differences in polarity preferences. Therefore, before computing geometric means across participants in the present experiment, the responses for each participant were studied. Within a block, most participants applied a consistent mapping polarity (be it positive or negative) and made fairly monotonic responses, so that, for example, low frequencies were given lower numbers and higher frequencies were given higher numbers. However, some participants were not as consistent in their responses as the others were. This necessitated the creation of three polarity categories: "positive," "negative," and "no" polarity, as follows: For each listener in each block, the Pearson correlation coefficient was computed between the log of the responses and the log of the actual stimulus values. Data from a specific participant in a given block were considered to have "no" polarity, and were not used in subsequent slope analyses, if the absolute value of the correlation coefficient in that block did not reach conventional levels of statistical significance ($r_{\text{critical}} = .444$, $p < .05$). That is, there was no significant linear component to the relationship between the sound parameter and the data it was supposed to represent. Note that this is a fairly generous limit. M. A. Teghtsoonian (1980, p. 296) has used much more stringent requirements, excluding data which do not achieve an r^2 of .70 (i.e., 70% of the variance in log judgment accounted for by variation in log stimulus value), which corresponds to $r = .84$. Data for which the correlation coefficient reached statistical significance were categorized as having positive or negative polarities, depending on the sign of the correlation, and were included in subsequent analyses. Table 2 includes the number of participants whose data resulted in positive, negative, and no polarities in each data-to-display block type.

The individual data from each participant within each block were also investigated for any evidence that demographic variables

Table 2
Summary of Results From Experiment 2

Display dimension	Slope of regression line, inside 95% confidence interval						
	Temperature	Pressure	Velocity	Size	Dollars	Pitch	Tempo
Frequency							
Pos. polarity	$0.68 \leq \mathbf{0.95} \leq 1.22$ ($n = 11$)	$0.66 \leq \mathbf{0.78} \leq 0.89$ ($n = 8$)	$0.99 \leq \mathbf{1.06} \leq 1.13$ ($n = 14$)	$0.78 \leq 0.90 \leq 1.02$ ($n = 7$)	$0.539 \leq \mathbf{0.77} \leq 1.00$ ($n = 6$)	$0.67 \leq \mathbf{0.78} \leq 0.89$ ($n = 16$)	
No polarity	($n = 3$)	($n = 4$)	(*none*)	($n = 1$)	($n = 5$)	($n = 1$)	
Neg. polarity	$-0.18 \leq -0.69 \leq -1.19$ ($n = 2$)	$-0.24 \leq -0.49 \leq -0.74$ ($n = 4$)	$-0.551 \leq -0.17 \leq +0.21$ ($n = 2$)	$-0.58 \leq \mathbf{-0.76} \leq -0.94$ ($n = 12$)	$-0.39 \leq -0.50 \leq -0.61$ ($n = 5$)	(*none*)	
Tempo							
Pos. polarity	$0.34 \leq \mathbf{0.43} \leq 0.52$ ($n = 11$)	$0.54 \leq \mathbf{0.68} \leq 0.82$ ($n = 10$)	$0.91 \leq \mathbf{1.04} \leq 1.19$ ($n = 11$)	(*none*)	$0.47 \leq \mathbf{0.66} \leq 0.85$ ($n = 8$)		$0.84 \leq \mathbf{0.95} \leq 1.06$ ($n = 13$)
No polarity	($n = 3$)	($n = 5$)	($n = 1$)	(*none*)	($n = 4$)		(*none*)
Neg. polarity	$-0.35 \leq -0.48 \leq -0.61$ ($n = 6$)	$-0.55 \leq -0.72 \leq -0.89$ ($n = 5$)	(*none*)	$-0.79 \leq \mathbf{-0.94} \leq -1.09$ ($n = 16$)	$-0.27 \leq -0.46 \leq -0.65$ ($n = 4$)	(*none*)	
Horizontal lines				$0.94 \leq \mathbf{0.98} \leq 1.02$ ($n = 16$)			

Note. The slope for the most popular polarity for each data-to-display pair is shown in boldface. No polarity was entered for a participant within a block if the correlation coefficient between the log of the responses and the log of the actual parameter values was less than .444. Pos. = positive; Neg. = negative.

might affect the mapping polarity or correlation coefficient. There were no significant correlations between any of the demographic variables (gender, handedness, and the number of years of music training) and the consistency of responding for any of the conditions.

Auditory Stimuli: Aggregate Analyses of Slope

For each polarity, geometric means were calculated for all judgments of a given stimulus across participants in a given data and display pair and were then plotted against the actual stimulus parameters. The slopes of all of the data-to-display mappings are summarized in Table 2, including both positive and negative slopes (where obtained), 95% confidence intervals, and the number of participants whose data contributed to each slope. If both polarities were obtained, the slope for the majority polarity is indicated in bold.

Perceptual dimensions: Pitch and perceived tempo. There were no negative polarity responses for either of the perceptual dimensions. The slope of the regression line for estimations of pitch versus actual frequency was $m = 0.78$ ($SE_m = 0.05$, $r^2 = .96$). The slope for perceived tempo versus actual tempo was $m = 0.95$ ($SE_m = 0.05$, $r^2 = .98$).

Conceptual dimensions. As indicated by the data in Table 2, within nearly all blocks there were some participants responding with positive and some with negative polarities. For example (see Table 2, second data column), the regression slope for pressure versus frequency resulting from the majority of listeners ($n = 8$) was $m = 0.78$ ($SE_m = 0.05$, $r^2 = .97$). The corresponding slope for listeners ($n = 4$) responding with a negative polarity is $m = -0.49$ ($SE_m = 0.08$, $r^2 = .83$). This example reflects twice as many listeners responding with the positive polarity as those responding with the negative polarity. For some mappings, the majority, or even unanimous response pattern was in the negative polarity. For example, all participants in the size:tempo block responded that increasing tempo corresponds to decreasing size.

Discussion

Horizontal Lines: A Replication

The data for the participants who saw horizontal lines and estimated their lengths yielded a complete replication of both the previous experiment and other findings in the literature. This confirmed that for line length, both the preferred polarity and the slopes within a block are somewhat stable across groups of participants drawn from the same population.

Auditory Stimuli

When participants listened to sounds that varied in frequency or in tempo, the perceptual estimations of pitch or tempo followed the patterns expected from related studies in the literature, both in terms of polarity and slope. This provides a baseline for the relationship between these two acoustic parameters and their perceived sensations. However, when groups of listeners heard sounds that varied in frequency or in tempo and were asked to make judgments about how much change in the conceptual data dimension a given change in the sound dimension represented, both the preferred polarity and the slope value depended on the data dimension.

The proportion of listeners who responded in each polarity within a block should begin to provide a measure of how effective a sonification mapping might be. This may serve as a design guideline for sonifications. That is, more agreement in polarities should be preferred over less agreement. For example, in the dollars:frequency mapping, only 6 of 16 participants favored a positive polarity (5 participants favored the negative polarity, and 5 yielded the no polarity). This seems to reflect a less intuitive pairing than, say, the temperature:frequency mapping that yielded 11 of 16 listeners in favor of a positive mapping (2 favored the negative mapping, and 3 yielded the no polarity preference). A mapping with unanimous, or near-unanimous support for a given polarity could be considered a good mapping, likely leading to fewer confusions when used in a sonification. An example would be the velocity:frequency mapping, where 14 of 16 participants responded with a positive polarity, versus two negative polarity responses. The size:tempo mapping showed that a negative polarity can certainly also be unanimously preferred for a mapping, even if the positive polarity is preferred for other mappings that make use of the same sound attribute.

Conceptual Models and Polarities

Comments from some participants about mental models they used while listening indicated a cognitive translation was definitely occurring between the sounds and the meanings being attached to the sounds. Participants were listening thoughtfully and not just uniformly applying the same mapping polarity and ignoring the content of the dimensions. The cognitive translation was not rigid, and in many cases listeners switched from a positive polarity to a negative polarity, and vice versa, from block to block.

Preference for both positive and negative polarities by different participants within the same mapping can result when the listeners have different mental models. This aspect has been examined more closely in a different study that compared magnitude estimation by sighted and visually impaired listeners (Walker & Lane, 2001). In one example from that study that is applicable to the present experiment, when asked about the dollars:frequency mapping, many of the sighted listeners described using an abstract higher: more model, reminiscent of most visual graphs. Number of dollars was treated as an abstract dimension, and a positive polarity mapping was preferred. The visually impaired listeners, however, reported using a more reality-based model, where dropping a small stack of dollar bills on a table sounds like a tap, a larger stack makes a lower frequency plop, and a bag full of bills makes a deep thud. In this model, dollars are dealt with on a more tangible level, resulting in a different mapping polarity.

Clearly, "explanations" of polarity choice are often somewhat speculative, even if provided by the participants themselves. Nevertheless, it points out that split polarities may mean qualitative differences in the conception of a sonification. Designers may be able to use magnitude estimation results to predetermine where a particular mental model (e.g., more money:lower frequency) intended by the designer needs to be explicitly encouraged in listeners.

Scaling Functions

As anticipated, based on the results of Experiment 1, the actual slope values for the scaling functions depended on the data dimen-

sion in play. Groups of listeners who heard exactly the same sounds produced different magnitude estimation slopes, depending only on the conceptual name provided for the data dimension. The slopes generally differ from each other and from the slopes obtained for simple perceptions of the acoustic parameters (see confidence intervals in Table 2). This suggests that the most effective way to use an auditory display parameter to represent changes in a data dimension is to take into account the slope of that scaling equation, in addition to polarity preferences.

Experiment 3: Validation of Slopes

The previous experiment showed that preferences, polarities, and the scaling function for the mapping of a data dimension to an auditory display parameter all depend on the exact data dimension in play. An open question arising from these findings is whether the use of unique polarities and slopes for each data and display pairing (as opposed to, say, a uniform linear data-to-display mapping with a slope of +1.0) actually results in improved performance with a sonification system in a real-task environment. The first stage in assessing this is to validate the stability and representativeness of the polarity preferences and slopes determined with the magnitude estimation procedure, as in Experiment 2. One obvious form of validation is replication. Asking many different groups of listeners to make magnitude estimations about different sonified conceptual data dimensions will converge toward an estimation of population preferences for both polarities and scaling slopes.

Another complementary approach is to assess how well the results of magnitude estimation with one group of participants are accepted as "natural" or "correct" by another similar set of listeners. The goal of Experiment 3 was to take steps in that direction, beginning to validate sonification polarities and slopes from Experiment 2 in a somewhat more tasklike environment. Many sonification tasks require the listener to judge the amount of change in a data dimension, as represented by a change in a sound parameter (e.g., Childs, 2001; McCabe & Rangwalla, 1994). In light of that, this experiment presented simplified auditory graphs and asked listeners to indicate which of two sets of data values each sound pattern best represented. One of the sets of data values was computed by using a power function, with the slope determined in Experiment 2, whereas the other set of values was computed by using a different slope value. If the sonification slopes determined in Experiment 2 are used in this sort of a data-analysis task, and if those slopes are preferred over both shallower and steeper slopes, then that should provide evidence converging toward the validation of the slope values. An additional goal in this experiment was to confirm that the level of agreement about a preferred polarity (from Experiment 2) predicts success on a subsequent sonification-interpretation task.

Method

Participants

A total of 60 students completed the experiment (28 men and 32 women; mean age = 20.1 years). Of these participants, 12 received $5 for participating rather than course credit.

Sound Stimuli: "Auditory Graphs"

There were two sets of stimuli in the form of simple auditory graphs, one set based on frequency (F) and the other set based on tempo (T). The three stimuli (i.e., auditory graphs) in the frequency set were each made up of a series of five 1-s pure tones separated by 0.25 s of silence. Each stimulus in this set sounded like a slow arpeggio played on an unusual scale. Stimulus F1 had frequency steps of 200, 250, 300, 350, and 400 Hz. Stimulus F2 had frequency steps of 200, 300, 400, 500, and 600 Hz. Stimulus F3 had frequency steps of 200, 400, 600, 800, and 1000 Hz. The amplitude envelope of the tones in this set of stimuli included a 0.1-s linear ramp onset (attack) and offset (release), and each of the steps was scaled for equal loudness to match an 800 Hz tone at 60 dBA SPL.

The three stimuli in the tempo set were also composed of five steps each. The steps in each of these stimuli increased in tempo rather than in frequency. The steps were composed of a repeating pattern of 0.200 beat of sound and 0.050 beat of silence (an on–off pattern, as in Experiment 2). The pattern was repeated at a certain tempo (e.g., 60 bpm) for as many repetitions as were required to fill approximately 1 s per step. The next steps were composed in the same manner, but the on–off pattern was repeated at progressively faster tempos. These subsequent steps were appended directly to the end of the previous step, for a total of five steps. Specifically, Stimulus T1 had tempo steps of 60, 75, 90, 105, and 120 bpm. Stimulus T2 had tempo steps of 60, 90, 120, 150, and 180 bpm. Stimulus T3 had tempo steps of 60, 120, 180, 240, and 300 bpm. Because of the fast repetition rates and brief on–off patterns, the amplitude envelope of the "on" part of the sound (the "tones") in this set of stimuli included a 0.01-beat linear ramp onset and offset, rather than the 0.1-beat ramps of the other stimuli. The frequency of the sine-wave tone components was 800 Hz for all steps of each stimulus in the tempo set. Further, the stimuli in the tempo set were all synthesized at the same loudness as the 800-Hz component of the frequency set, giving all of the stimuli in both sets equal perceived loudness.

Word Stimuli: "Data Patterns"

The task in this experiment was to listen to an auditory graph (i.e., one of the sound stimuli described above), then determine which of two sets of numerical values ("data patterns") that auditory graph best represented. The data patterns were created as follows:

Consider the example where Stimulus F1 was played and the listener had to make a judgment about pressure. The starting frequency of the stimulus (f_1 = 200 Hz) was defined as being equal to an initial data (i.e., pressure) value of P_1 = 100 units. The final frequency of the stimulus was, in this case, f_5 = 400 Hz. To calculate the final pressure, P_5, the equation $P_5 = P_1 \times (f_5/f_1)^m$ was used, where m is the slope from the regression equation determined in the previous experiment. The regression slope for estimated pressure P versus actual frequency f, obtained in Experiment 2, was m = 0.7812. Substituting the values from the present example yields P_5 = 172. Thus, for the Stimulus F1 (starting at 200 Hz and ending at 400 Hz), the calculated data values for pressure started at 100 units of pressure and ended at 172 units of pressure.

To make the task somewhat easier for the participant, the starting value of each data dimension was 100 on every trial, for all participants. Also, as previously described, the starting value of pitch or tempo was constant within a block of trials. To reduce the number of experimental conditions, only positive slopes were used here, even if there were both positive and negative slopes obtained in Experiment 2. In the case of the size:tempo mapping, every participant in that block in Experiment 2 had responded with a negative slope, so there was no positive slope to use. Therefore, in the present experiment the slope used in that block was a somewhat arbitrary +1.0.

For each "correct" data pattern, calculated as described above, two "incorrect" data patterns were created by multiplying the correct data endpoint by 0.80 and 1.20, respectively. Thus, if the correct data pattern went from 100 to 172 units, the incorrect patterns would be from 100 to 138, and from 100 to 206 units (80% and 120% of the endpoint, respectively).

Design

The design again included the factors of Display Dimension (frequency and tempo) and Data Dimension (temperature, pressure, velocity, size, and number of dollars). There were no perceptual data dimensions (pitch or perceived tempo) included in this experiment. Each data dimension (e.g., temperature) was paired with one of the display dimensions (e.g., frequency) for an entire block of trials (e.g., temperature:frequency). Each participant again completed two blocks of trials separated by a brief rest. Twelve participants completed each block type.

Trial Structure and Task

In the top center of the screen, there was a 2.5-cm square graphic with the words, "START HERE." Below the sound square were two boxes, side-by-side on the screen. Each contained the data dimension and the data pattern endpoints; for example, "Pressure starting at 100, and ending at 172." This example includes 172 as the endpoint, which for pressure would be a correct trial, as calculated above. On any given trial, the correct data pattern could be presented in one of the boxes, with either of the corresponding incorrect data patterns in the other box, or else both of the incorrect patterns could be presented as a foil trial.

The participant moved the cursor over the square to play the stimulus. On each trial, the listener heard one sound stimulus from the set being used for that block (e.g., the frequency set) and decided which of the two data patterns shown on the screen the auditory graph best represented. Participants clicked a button underneath the data pattern that they felt matched the sound pattern. This was a forced-choice design, requiring a response even when there were two incorrect data patterns, though the participants had no knowledge that any of the data patterns were considered correct or incorrect.

In a block of 18 trials, each of the three sound stimuli in a set was played twice with every possible pairing of its corresponding correct and incorrect data patterns. The location of the data patterns was counterbalanced left-to-right, and all of the trials were presented in a randomized order.

Results

Experimental Trials

The data from the two display dimensions (frequency and tempo) were separated and then sorted by data dimension and participant. For the 12 trials where the correct data pattern had been present, each response was scored based on whether the listener had picked the correct data pattern. The number of correct responses divided by 12 determined the proportion correct for that participant. The overall grand mean proportion correct was calculated across all participants and all blocks within each display dimension, and showed that participants picked the correct data pattern in this experiment significantly more often than would be expected by chance. These results form the leftmost column of data in Table 3. For frequency, the mean proportion correct of 0.61 ($SD = 0.14$) was significantly higher than 0.50, $t(59) = 5.82$, $p < .0001$. Similarly, when tempo was the display dimension participants also picked the correct data pattern significantly more often than would be expected by chance (mean proportion correct = 0.5958, $SD = 0.1254$), $t(59) = 5.92$, $p < .0001$.

Next, the grand mean proportion correct was calculated across all participants within each separate data type to determine whether the participants in that block had, as a group, responded better than chance. Table 3 (Data Columns 2–6) summarizes the grand mean proportion correct for each of the trial block types when frequency was the display dimension and when tempo was the display dimension. As seen in the table, the mean proportion correct for each block type was numerically larger than 0.50. With frequency as the display dimension, this better-than-even preference reached statistical significance for temperature, velocity, and size, but not for pressure or dollars. With tempo, all of the data dimensions except velocity reached statistical significance.

Foil Trials

Preference was also analyzed for the trials where neither of the data patterns was correct. Performance on these trials should be statistically equivalent to guessing. For each block type (i.e., each mapping type), the overall proportion of foil trials where participants chose the lower of the two incorrect data patterns was calculated. This score was a measure of whether one or the other of the incorrect data patterns had been chosen more often than would be expected by chance. Ten separate t tests showed that in none of the block types was the mean proportion statistically different from 0.50 for any of the block types. Similarly, the overall grand mean proportion of foil trials where the lower pattern was chosen, across block types, was also calculated for each of the two stimulus sets. The overall grand mean for foil trials in the frequency set, 0.49, was not significantly different from 0.50, $t(14) = 0.27$, $p = .79$. For foil trials in the tempo set, the overall grand mean, 0.43, was again not significantly different from 0.50, $t(14) = 1.70$, $p = .11$. Thus, the responses for the foil trials with

Table 3
Grand Mean Proportion Correct in Each Block Type in Experiment 3

Display	All data ($N = 60$)	Temperature ($n = 12$)	Pressure ($n = 12$)	Velocity ($n = 12$)	Size ($n = 12$)	Dollars ($n = 12$)
Frequency set						
Proportion correct	0.607	0.639	0.549	0.639	0.653	0.556
Variance	0.020	0.012	0.022	0.017	0.031	0.016
t	5.82**	4.43**	1.13	3.71**	2.99*	1.54
p	.0001	.0010	.2808	.0035	.0123	.1513
Tempo set						
Proportion correct	0.596	0.597	0.611	0.514	0.660	0.597
Variance	0.016	0.011	0.019	0.014	0.021	0.007
t	5.92**	3.19**	2.77*	0.41	3.84**	3.92**
p	.0001	.0086	.0183	.6887	.0028	.0024

* $p < .05$. ** $p < .01$, two-tailed.

both the frequency and tempo sets of stimuli indicated no systematic preference for one or the other of the incorrect data patterns.

Discussion

Slope Validation

Participants heard simple auditory graphs and for each one judged which of two data patterns the auditory graph represented. Overall, listeners preferred the data pattern that used the experimentally determined slopes. This provides evidence converging toward validation of the slopes obtained in Experiment 2. Whereas in all cases it was numerically better than chance, in a few of the specific data-to-display pairings the preference for the correct data pattern did not reach conventional levels of statistical significance. Specifically, the pressure:frequency, dollars:frequency, and velocity:tempo sets did not result in a preference score that was significantly different from guessing.

The practical implications resulting from this particular validation experiment are necessarily limited by the fact that it uses a new approach to assessing how preferences obtained with one group (via magnitude estimation) match the preferences of another group. As with any new technique, the validation method used here would need to be used in other contexts and studied further to confirm its utility, as the data here are considered in conjunction with other sources of converging evidence. The comments of several of the participants who indicated that they had chosen the patterns "that had seemed right" add to the conclusion that this approach is diagnostic of the participants' actual preferences.

Mapping and Polarity Validation

The results of Experiment 3 also provide support for the use of the ratio of the number of responses in a particular polarity to the number of all responses as a predictor of the "naturalness" of the mapping. A simple "overall-majority" rule was applied here: If a given polarity obtained a majority of all responses by participants in a block in Experiment 2, it was predicted to be a "good" or natural-polarity choice. Less than a majority resulted in the prediction of a "bad," or unnatural-mapping choice. Table 4 summarizes the predictions that emerge from the results of Experiment 2. Question marks represent uncertain mappings. For example, in the case of temperature:frequency, the positive polarity had 11 of 16 responses in Experiment 2, predicting that it would be a good polarity. This was supported in the results of Experiment 3 (see Table 3). However, the positive polarity in the dollars:frequency mapping had only 6 of 16 responses in Experiment 2, resulting in a bad prediction. Again, this prediction was supported by the results in Experiment 3. The positive polarity for the pressure: frequency mapping had 8 of 16, or half of the total responses, making it an ambiguous-mapping choice. In Experiment 3 that mapping did not lead to a significant preference score. Applying the decision heuristic, the positive polarities in the pressure: tempo and dollars:tempo mappings would also be uncertain; the former resulted in good performance, whereas the latter resulted in poor performance. This demonstrates the importance of identifying ambiguous mappings that can lead to confusions in sonifications.

Table 4
Predicted Effectiveness of Frequency and Tempo for Displaying Various Conceptual Display Dimensions in Experiment 3, Based on Overall Majority of Polarities in Experiment 2

	Data dimension				
Display dimension	Temperature	Pressure	Velocity	Size	Dollars
Frequency					
Positive polarity	Good (73%)	Uncertain (50%)	Good (88%)	Poor* (35%)	Poor (38%)
Negative polarity	Poor (13%)	Poor (25%)	Poor (12%)	Good (60%)	Poor (31%)
Tempo					
Positive polarity	Good (55%)	Uncertain (50%)	Good* (92%)	Poor* (0%)	Uncertain (50%)
Negative polarity	Poor (30%)	Poor (25%)	Poor (0%)	Good (100%)	Poor (25%)

Note. The descriptors represent only the predicted effectiveness, based on the overall level of unanimity of the mapping polarity, replicating Walker and Kramer (1996). The percentages of participants responding with each polarity are provided in parentheses. Asterisks indicate an incorrect prediction.

The predictions were correct for 7 of the 10 mappings used in Experiment 3 (in Table 4, the three incorrect predictions are denoted by asterisks). Two of the erroneous predictions might also be tempered by the fact that they involved the size data dimension. There was majority support for the negative polarities in those mappings in Experiment 2, but Experiment 3 only included the positive polarity. In one of the cases, listeners performed better than predicted; in the other case, they performed worse than predicted. It is unclear why the velocity:tempo mapping, predicted to be good on the basis of the results of Experiment 2, did not result in better-than-even preference in Experiment 3.

In sum, the present experiment provided evidence validating the results of Experiment 2 by providing initial confirmation of the slope values, and also by supporting the use of the level of overall support for a polarity as a predictor of the effectiveness of a mapping when used in a subsequent sonification. As stated, all three categories of polarities need to be considered in this heuristic.

General Discussion

Despite increasing popularity and the growing number of scientific success stories for sonification, there has been very little experimental evaluation of how to construct such auditory graphs. In particular, there has been no systematic evaluation of the way data values are mapped onto auditory display values. The research presented here provides a start at answering three main questions that such an evaluation needs to address: (a) What is the best sound parameter to use to represent a given data type? (b) Should an increase in the sound dimension (e.g., rising frequency) represent an increase or a decrease in the data dimension (e.g., temperature)? (c) How much change in the sound dimension would represent a given change in the data dimension?

When participants in the present experiments made conceptual estimations about the data values (e.g., temperature, pressure) that the display dimensions would represent, the slopes were nearly all different from each other, and, importantly, different from the slopes obtained for the perceptual dimensions. This applied both to

visual display dimensions (e.g., horizontal lines) and auditory display dimensions (e.g., frequency). Participants described using mental models of physical systems, showing that there is a cognitive translation involved in the mapping of a conceptual data dimension to a display dimension. This resulted in both positive and negative polarities being preferred, depending on the mapping. Clearly, the most successful representation of conceptual data depends on the most appropriate display dimension being used, and in the right way.

Sonification design is not a simple task in any case. The finding that it matters what sound attribute is used to represent some data makes it even more challenging. However, the results here indicate that listener preferences for both mapping polarities and psychophysical scaling functions can be determined simply and effectively by using magnitude estimation.

Researchers have shown in many studies that poorly designed visual displays can have serious effects on performance and that improvements in those displays can have measurable performance gains (e.g., Sanders & McCormick, 1993). Experiment 2 points to the same conclusion for auditory displays, consistent with the performance results from Walker and Kramer (1996). Walker and Ehrenstein (2000) have shown performance degradation with auditory stimuli as a result of stimulus-response compatibility effects, which can be considered an interaction between the auditory mapping characteristics and the listener expectations. Ignoring the preferred mappings for conceptual dimensions is likely to lead to similar conflicts in the perceive-think-respond action chain involved in the use of auditory displays and sonifications (see also, Sorkin, 1988).

Validation of Preferences

Polarity preferences and scaling functions should be validated for a given listener group to ensure maximal agreement between the sound design and the listeners' expectations. This would ensure optimal performance with the resulting sonification. One obvious means of validation—replication—was supported here by the fact that the two magnitude estimation experiments replicated findings in the literature for similar perceptual dimensions, while finding different values for the conceptual dimensions. Also, visual participants in Experiment 2 replicated the findings from Experiment 1. Further replication with new listeners and different variations of the display parameters would provide additional support for other conceptual dimensions. This approach is currently underway in a separate line of research (Walker, 2002).

It is desirable to approach validation from multiple angles, not only through replication. Experiment 3 in this project served as the initial step in a novel approach to validating both the polarity preferences and the scaling factors. Listeners in Experiment 3 preferred the mappings that used experimentally determined scaling slopes. In addition to this slope preference, the findings supported the use of a polarity's overall unanimity (from magnitude estimation) as a measure of the naturalness, or likely effectiveness of a mapping for displaying a given data dimension.

It is also instructive to compare the predictions and results in this project with previous findings, such as those presented by Walker and Kramer (1996; also described in Walker, 2000). Those recommendations were based on a combination of accuracy and reaction time results with a different task, and included fewer data dimensions, more display dimensions, and only positive polarities. Even with the differences in the studies, the results from Walker and Kramer (1996) are in general agreement with the present findings. The present study generally led to more definitive predictions (Walker and Kramer called 8 of their 16 mappings "okay"—neither good nor bad, overall). There were, though, three mappings in the present study for which Walker and Kramer also made definitive assessments. They called velocity:pitch (positive polarity) a "good" mapping; this was supported by both the prediction rule and the performance in Experiment 3 of the present study. Walker and Kramer called size:tempo (positive) a "bad" mapping. That is the same as what was predicted here, based on the results of Experiment 2. Curiously, performance on size:tempo (positive) in Experiment 3 here contradicted both our predictions and the assessments by Walker and Kramer. Finally, Walker and Kramer called pressure:tempo (positive) a "bad" mapping. The present study had an agnostic prediction based on Experiment 2, and the actual performance in Experiment 3 was good. Overall, the present study's prediction heuristic seems to provide a fairly specific and successful method for evaluating mapping choices in advance and aligns with recommendations made in previous studies.

Magnitude Estimation in Sonification Design

The fact that sonification designers have had little theory and few guidelines is compounded by the realization that what sounds "intuitive" to one, or several designers may not match the conceptions of the intended audience (e.g., Walker & Kramer, 1996). When preference data are not available or not applicable to the particular needs of a sonification project, using magnitude estimation as part of the initial design phase is recommended as a simple and useful way to learn about the expectations held by listeners. It highlights cases where the listeners' mental models differ from that of the display designer in significant ways. The information about polarities can help a designer predict which mappings would be successful and can either allow for a redesign or pointing out potential areas where instructions, training, or other methods may be necessary. Magnitude estimation also provides the necessary slope values relating data and display dimensions. This recommendation echoes the efforts of Edworthy et al. (1995) in using magnitude estimation results to tailor the perceived urgency of cockpit warning sounds. In the case of displaying data, as investigated here, sonification toolkits would need to allow for different scaling functions for each data and display pair; these functions also need to be flexible enough to incorporate the results of further research.

Continuing Research Needs

Many more acoustic parameters, resulting in many more sets of sounds, would need to be tested before widely generalizable sonification guidelines can be developed. Also, as much as Hellier et al. (1993, 1995) have begun to consider more complex sounds in the perception of warnings, data sonifications also need to be studied in more dynamic forms. It is also important to consider the perceptual reactions from a more diverse group of listeners. This seems especially likely if the groups differ widely in the nature of their particular listening experience. For example, recent research (e.g., Walker & Lane, 2001) has used magnitude estimation to consider the auditory expectations of visually impaired listeners.

Interesting differences in the patterns of results between sighted and visually impaired participants indicate that listening experience does affect mapping preferences. Application developers may need to apply this paradigm to their specific target audience to catch any such variability.

The final test would always be instantiating these and other findings in more and varied sonification applications and systematically evaluating their effectiveness. With a better handle on the mappings, scalings, and auditory graphing techniques, we can continue to implement sonifications and auditory graphs that have greater practical utility for researchers, teachers, and students, both sighted and visually impaired, in all manner of scientific disciplines.

References

Barrass, S. (1998). *Auditory information design*. Unpublished doctoral dissertation, Commonwealth Scientific and Industrial Research Organisation, Australia.

Beck, J., & Shaw, W. A. (1961). The scaling of pitch by the method of magnitude estimation. *American Journal of Psychology, 74,* 242–251.

Beck, J., & Shaw, W. A. (1962). Magnitude estimations of pitch. *Journal of the Acoustical Society of America, 34,* 92–98.

Beck, J., & Shaw, W. A. (1963). Single estimates of pitch magnitude. *Journal of the Acoustical Society of America, 35,* 1722–1724.

Childs, E. (2001). The sonification of numerical fluid flow simulations. In J. Hiipakka, N. Zacharov, & T. Takala (Eds.), *Proceedings of the Seventh International Conference on Auditory Display, ICAD 2001* (pp. 44–49). Espoo, Finland: ICAD.

Dawson, W. E., & Brinker, R. P. (1971). Validation of ratio scales of opinion by multimodality matching. *Perception and Psychophysics, 9,* 413–417.

Edworthy, J., Hellier, E., & Hards, R. (1995). The semantic associations of acoustic parameters commonly used in the design of auditory information and warning signals. *Ergonomics, 38,* 2341–2361.

Edworthy, J., Loxley, S., & Dennis, I. (1991). Improving auditory warning design: Relationship between warning sound parameters and perceived urgency. *Human Factors, 33,* 205–232.

Eisler, H. (1976). Experiments on subjective duration 1868–1975: A collection of power function exponents. *Psychological Bulletin, 83,* 1154–1171.

Ekman, G. (1958). Two generalized ratio scaling methods. *Journal of Psychology, 45,* 287–295.

Engen, T. (1971). Psychophysics II: Scaling methods. In J. W. Kling & L. A. Riggs (Eds.), *Experimental psychology* (3rd ed., pp. 47–86). London: Methuen.

Hellier, E. J., Edworthy, J., & Dennis, I. (1993). Improving auditory warning design: Quantifying and predicting the effects of different warning parameters on perceived urgency. *Human Factors, 35,* 693–706.

Hellier, E., Edworthy, J., & Dennis, I. (1995). A comparison of different techniques for scaling perceived urgency. *Ergonomics, 38,* 659–670.

Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J., Miner, N., Neuhoff, J., Bargar, R., Barrass, S., Berger, J., Evreinov, G., Fitch, W., Gröhn, M., Handel, S., Kaper, H., Levkowitz, H., Lodha, S., Shinn-Cunningham, B., Simoni, M., & Tipei, S. (1999). *The sonification report: Status of the field and research agenda*. Report prepared for the National Science Foundation by members of the International Community for Auditory Display. Santa Fe, NM: International Community for Auditory Display.

Lane, H. L., Catania, A. C., & Stevens, S. S. (1961). Voice level: Autophonic scale, perceived loudness, and effects of sidetone. *Journal of the Acoustical Society of America, 33,* 160–167.

McCabe, K., & Rangwalla, A. (1994). Auditory display of computational fluid dynamics data. In G. Kramer (Ed.), *Auditory display: Sonification, audification, and auditory interfaces* (pp. 327–340). Reading, MA: Addison Wesley.

Sanders, M. S., & McCormick, E. J. (1993). *Human factors in engineering and design* (7th ed.). New York: McGraw-Hill.

Sorkin, R. D. (1988). Why are people turning off our alarms? *Journal of the Acoustical Society of America, 84,* 1107–1108.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley.

Stevens, S. S., & Guirao, M. (1963). Subjective scaling of length and area and the matching of length to loudness and brightness. *Journal of Experimental Psychology, 66,* 177–186.

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *American Journal of Psychology, 53,* 329–353.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America, 8,* 185–190.

Teghtsoonian, M. A. (1965). The judgment of size. *American Journal of Psychology, 78,* 392–402.

Teghtsoonian, M. A. (1980). Children's scales of length and loudness: A developmental application of cross-modal matching. *Journal of Experimental Child Psychology, 30,* 290–307.

Teghtsoonian, M., & Teghtsoonian, R. (1971). How repeatable are Stevens's power law exponents for individual subjects? *Perception and Psychophysics, 10,* 147–149.

Teghtsoonian, M., & Teghtsoonian, R. (1983). Consistency of individual exponents in cross-modal matching. *Perception and Psychophysics, 33,* 203–214.

Walker, B. N. (2000). *Magnitude estimation of conceptual data dimensions for use in sonification*. Unpublished doctoral dissertation, Rice University, Houston, TX.

Walker, B. N. (2002). Replication as validation of sonification preferences. Manuscript in preparation.

Walker, B. N., & Ehrenstein, A. (2000). Pitch and pitch change interact in auditory displays. *Journal of Experimental Psychology: Applied, 6,* 15–30.

Walker, B. N., & Kramer, G. (1996). Mappings and metaphors in auditory displays: An experimental assessment. In S. Frysinger & G. Kramer (Eds.), *Proceedings of the Third International Conference on Auditory Display, ICAD '96* (pp. 71–74). Palo Alto, CA: ICAD.

Walker, B. N., & Lane, D. M. (2001). Psychophysical scaling of sonification mappings: A comparison of visually impaired and sighted listeners. In J. Hiipakka, N. Zacharov, & T. Takala (Eds.), *Proceedings of the Seventh International Conference on Auditory Display, ICAD 2001* (pp. 90–94). Espoo, Finland: ICAD.

Williams, R. L. (1956). *Statistical symbols for maps: Their design and relative values*. New Haven, CT: Yale University, Map Laboratory.

Received May 7, 2002
Revision received May 9, 2002
Accepted May 9, 2002 ■