The Sonification Handbook

Edited by
Thomas Hermann, Andy Hunt, John G. Neuhoff

Logos Publishing House, Berlin, Germany
2011, 586 pages

Online: http://sonification.de/handbook
Order: http://www.logos-verlag.com

Reference:

Chapter 6

Evaluation of Auditory Display

Terri L. Bonebright and John H. Flowers

Reference:

Media examples: http://sonification.de/handbook/chapters/chapter6
Chapter 6

Evaluation of Auditory Display

Terri L. Bonebright and John H. Flowers

6.1 Chapter Overview

Evaluation of auditory displays is a crucial part of their design and implementation. Thus, the overriding purpose of this chapter is to provide novice researchers with basic information about research techniques appropriate for evaluating sound applications in addition to providing experienced perceptual researchers with examples of advanced techniques that they may wish to add to their toolkits. In this chapter, information is presented about general experimental procedures, data collection methods for evaluating perceptual qualities and relations among auditory stimuli, analysis techniques for quantitative data and distance data, and techniques for usability and active user testing. In perusing the information in this chapter, the reader is strongly urged to keep the following issues in mind.

First and foremost, all application development should have ongoing investigation of the perceptual aspects of the design from the beginning of the project (Salvendy, 1997; Sanders & McCormick, 1993; Schneiderman, 1998). It is an extreme waste of time and other resources to finish an auditory display and then have the target audience attempt to use it. Such mistakes in the design process used to be common in computer design, but the work of such individuals as Schneiderman (1998) in human-computer interaction work has made ongoing evaluation a regular part of computer software and hardware development for most companies. The same approach should be used for auditory display design as well.

Second, the reader should note that the choice of research method has to be intimately tied to the final goal of the project. If the project is designed to develop a full-scale sonification package for a specific group (for examples of such projects, see Childs, 2005; Valenzuela, Sansalone, Krumhansl, & Street, 1997), the researcher would need to use a variety of methods including both laboratory components and ecologically valid testing. For example, it might be shown in the laboratory experiments that a specific sound for rising indexes of financial data works better than another; however the researcher might find in the real-world
application with stockbrokers that the winning sound from the lab may have spectral overlap with noises in the environment and therefore won’t work for the target application. In this case, it would be ideal if sounds in the target environment were specified during an analysis and specification phase for the project prior to testing in the laboratory.

Third, it is important to note that even though this chapter extols the virtues of research techniques for developing good audio applications, experts should also use their own introspection and intuition, especially when beginning a project. Such expertise can be tremendously useful in narrowing down what might otherwise be a Herculean task for something as simple as determining which sounds might be most appropriate. But researchers should not rely on their expertise alone and must do actual testing to determine how well the application will work for the target audience. This can be illustrated most clearly when an expert in visualization techniques assumes that auditory displays can be developed using the same principles. Unfortunately, there are different perceptual properties that come to bear on building effective auditory display applications, such as limitations of sensory and short-term memory, that are less relevant to the design of visual displays.

As a final introductory point, the reader should be aware that it is not the intent of this chapter to replace any of the excellent references that are available on research design issues or data analysis techniques. Rather, the purpose is to provide an overview of the research process as it pertains specifically to auditory display design. Embedded within this overview are referrals to more detailed and in-depth work on each of the relevant topics. It is also hoped that this chapter will foster interdisciplinary collaboration among individuals who have expertise in each of the disciplines that contribute to auditory display design, such as cognitive and perceptual psychologists, psychoacousticians, musicians, computer scientists, and engineers, since this leads to the most rapid development of good applications.

### 6.2 General Experimental Procedures

In this section general information is presented about design issues pertinent to the investigation of perceptual characteristics of auditory stimuli. The first issue in designing an empirical study for sound applications is to have a clear idea of the goals for the specific auditory display of interest, which are then used to develop the questions to be considered in the study. It is important to emphasize that all experimental procedures must be developed within the context of the particular application and setting. Thus, each issue discussed in this section assumes that the context and the goal for the application are embedded within each decision step for setting up the study.

The second issue a researcher needs to consider is what types of data and statistical analyses are required to answer the questions of interest. One of the major problems experienced by novice researchers is that they fail to recognize that it is critical that data analysis techniques must be specified during the design stage, since they impact directly the type of data that should be collected, as well as other design considerations discussed in this chapter.

The following material on general experimental procedures moves from overarching concerns (e.g. experimenter and participant bias), to basic design topics (e.g. number and order of stimuli) and finishes with participant issues (e.g. participant selection). Unfortunately, the actual process is not linear in nature but resembles a recursive loop, since the researcher
Evaluation of Auditory Display 113

needs to adjust design parameters in relation to each other in order to develop a successful procedure. (See Keppel & Wickens, 2004 for a good general reference for research design for behavioral studies.)

6.2.1 Experimenter and Participant Bias

Experimenter effects occur when the investigators collecting the data either treat participants in experimental conditions differently or record data in a biased manner. Typically such bias happens when the experimenter has expectations about the probable or “desired” outcomes of the study and inadvertently impacts the participants in such a way that it modifies their responses. This is an especially crucial issue during usability and active use testing or when an investigator is conducting any type of interview procedure. It is noteworthy that investigators who are in a power hierarchy, such as graduate or undergraduate research assistants working with a professor, may be more prone to the effects of experimenter bias in general. Supervisors should talk openly about such problems with their data collection team as part of the training process. This should help minimize the effects of any previous knowledge about the expected results investigators carry with them into the experimental sessions, as well as to alleviate any perceived pressure to “please” the authority figure.

Experimenter bias interacts with the tendency for participants in experiments to want to be “good subjects”, and as a consequence, they seek clues about what the “right” answer is, even if the investigator assures them that there is no such thing. Participants can be sensitive to these demand characteristics and provide feedback that reflects what they think the experimenter wants to have as the outcome. Obviously such bias on the part of both experimenters and participants is undesirable, and researchers can use a number of methods to reduce or eliminate these problems. For example, one common and quite effective practice for reducing demand characteristics is to have data collection performed by individuals who are “blind” to the hypotheses (and sometimes even the specific purposes) of the study. Another effective method is to automate the procedure as much as possible by using written or video recorded instructions and computerized testing.

6.2.2 Perceptual Limitations Relevant to Sound Perception

There are a number of cognitive and perceptual issues that are especially important for researchers interested in evaluating auditory displays. It is common for researchers new to the field to assume people’s processing capabilities for sounds are very similar to their abilities for visual stimuli. Unfortunately, some fundamental differences between auditory and visual perception make this a dangerous and misleading assumption. Discussions of many of these critical differences between hearing and vision can be found in Bregman (1990), Handel (1989), Hass and Edworthy (2002), and McAdams and Bigand (1993) - sources which researchers and developers should be encouraged to read. Three aspects of auditory perception that place constraints on tasks and methods used to evaluate auditory displays are the transient nature of sounds, properties of memory for auditory events, and differences in the way attention is allocated in auditory as opposed to visual tasks.

Since sounds exist in time and are transient, unlike static visual displays that can be repeatedly inspected and “re-sampled over time” at the will of the observer, re-inspection of sound
requires that it be replayed. Comparisons between sounds require that features of one sound be retained in memory while another is being heard, and/or that information about more than one sound be retained in memory at the same time. There are thus major limitations related to sensory memory, working memory, and long-term memory for sounds that are crucial to consider during the testing and design phases of a project. These limitations affect both the design of auditory display elements themselves, as well as how to go about effectively evaluating them. Specifically, these limitations constrain the optimum duration for a discrete auditory display presentation, the optimal duration between presentation of elements to be compared (the interstimulus interval\(^1\)), and the degree of control of the display that is given to a participant in an evaluation or a user study. For auditory display applications that present discrete “packages” of information by sound (e.g., earcons (see Chapter 14), auditory representations of discrete data samples, etc.) the designer usually has the ability to control display duration, and thus the determination of a duration that optimizes task performance should be one of the objectives of display evaluation. In designing and evaluating such applications participants or users will need to make comparisons between auditory displays (e.g., sorting tasks, similarity ratings). The effective duration of auditory sensory memory is an issue for making such comparisons; if displays or stimuli exceed 12 seconds or so, it is likely that memory for events at the beginning of the display will be degraded and the ability of participants to make reliable comparisons will be impaired. However, shortening the duration of a display of complex information runs the risk that perception of auditory patterns will be impaired because they are presented too rapidly. Thus there may be a three-way tradeoff between sensory memory, perception, and display complexity that designers need to consider and specifically investigate in designing such applications.

In most research designs, any task involving comparisons between auditory displays should be set up so that participants can repeat stimuli for as many times as they feel is necessary to make a good evaluation. The exception to this general rule is when the researcher desires to have an intuitive response, such as the almost reflexive response desired for an alarm; in such cases, the sounds should be limited to a single presentation. Additionally, if feasible, participants should be given control over the interstimulus interval, in order to ensure that there will be little interference between the perceptions of the stimuli. If it is necessary to have a fixed delay between display presentations, the interval should be long enough to allow perceptual separation between the displays, but not allow degradation of the sensory memory of the first display. A pilot study can be helpful to determine what seems to be a “comfortable” interstimulus interval for a given type of display - generally in the range of 0.5 to 4.0 seconds.

Evaluation of displays intended for on-line monitoring of continuous status information (e.g., industrial systems, patient vital signs in the operating room, etc.) present a somewhat different set of problems. The issue here is not the memory of the entire display but the detection of changes and patterns within the display which require action on the part of the observer. For these types of displays, most development research is concerned with determining optimal perceptual mappings between sound and data channels and how many streams of data to present (see Chapter 15). In such tasks, attention limitations are of particular importance, and these are generally assessed by measuring actual task performance by such measures as detection accuracy for “significant” events. However attentional capacity is also taxed

\(^1\) An interstimulus interval is the amount of time between the offset of one stimulus and the onset of the following stimulus.
significantly in most auditory testing situations, even those involving comparisons of, or decisions about, “short” discrete auditory displays; therefore, the researcher should take extra care to make sure that participant fatigue does not impact the quality of the resulting data (see Chapter 4 for information on both perceptual and cognition issues).

Ideally, researchers testing auditory displays would benefit greatly from having basic data about perceptual abilities, including limitations, for auditory display elements in a fashion similar to the data that have been compiled in anthropometry research (Dreyfus, 1967; Roebuck, Kroemer, & Thomson, 1975). These data provide measures of hundreds of physical features of people that are used by industry to provide means and percentile groupings for manufacturing most of the products people use that are related to body size. General information about auditory perceptual abilities is available in a variety of journal papers and other manuscripts (i.e. Bregman (1990), Hass & Edworthy (2002), Handel (1989), McAdams & Bigand (1993), and Salvendy, 1997), but not in one complete comprehensive compilation with the necessary norms for the populations of interest. Such a guide of auditory perceptual parameters for auditory display researchers would allow the development of sound applications for specific groups in addition to the construction of sound applications that could provide a range of sounds that would work for the majority of individuals within a heterogeneous population.

6.2.3 Number and Order of Stimuli

While designing a study, researchers need to determine the appropriate number of stimuli and how these stimuli will be presented to the participants. Researchers and developers should carefully consider the issues of working memory and cognitive load when deciding how many stimulus attributes will be manipulated (e.g., pitch, intensity, etc.) and how many levels or values will be varied per attribute. In cases for which the investigator wishes to study basic perceptual abilities (as might be the case in exploratory stages of auditory display development), it may be preferable to err on the side of fewer rather than more stimuli in order to obtain useful data. On the other hand, in later stages of display development, in which the goal is to evaluate a display design in a real-world environment, it may be necessary to manipulate all applicable variables to determine how the display will perform.

Repeated stimuli may be added to the total number of stimuli to test subject reliability. Typically, a small number of randomly selected stimuli are repeated and randomly placed in the stimulus order, so that participants are unaware of the repeated trials. Data from these repeat trials are then used for cross correlation coefficients\(^2\) to compute subject reliability. These correlation coefficients can then provide the researcher with information about which participants might be outliers since a low coefficient may indicate that the individual had a perceptual disability, did not take the task seriously, or did not understand the directions. Data from such participants are likely to provide an inaccurate picture of the perceptual response for the majority of the participants and lead to decisions about an auditory display that are misleading or incorrect.

Once the number of stimuli has been determined, the order of stimulus presentation should be considered. This is a particularly crucial issue for auditory stimuli since any stimulus

\[^2\text{More information on correlation can be found in section 6.4.1 and additional information about using such techniques for determining outliers in section 6.5.1.}\]
presented before another one has the possibility of changing the perception of the second stimulus. An example of this would be when a high amplitude, high frequency sound is presented directly before a sound that has low amplitude and low frequency. If the researcher is asking about basic information (such as perceived pitch or volume), the response for the second stimulus may be skewed from the exposure to the first sound. In studies where there are a very small number of stimuli (< 5), the best solution is to provide all possible orderings of stimuli. For most typical studies where the number of stimuli is too large to make all possible orders practical, the most effective method is to randomize the stimulus order, which distributes any order effects across participants and these are consequently averaged out in the composite data. Computer presentation allows for full randomization of stimuli across participants, but if stimuli must be presented in a fixed order (e.g., using pre-recorded audio media), then three or four randomly generated orders should be used.

### 6.2.4 Testing Conditions, Pilot Testing and Practice Trials

Decisions about the testing conditions under which data are collected should take into account the specific purpose of the study. For example, when conducting basic auditory perception research, it is essential to eliminate as many extraneous variables as possible (e.g., noise or visual stimuli) that could be distracting and to keep the environmental conditions constant across task conditions. On the other hand, for research projects designed to test the usability of a product for an industrial setting, the study should be conducted in the target environment. Regardless of the general testing conditions, instructions for the procedures should be carefully constructed and standardized in content and presentation for all participants.

The time it takes participants to complete an auditory display study is extremely important, since perceptual tasks tend to be demanding in terms of attention and vigilance, which can lead to participants becoming fatigued or losing motivation over the course of the session. As a general rule, most studies should have a limited task time of no more than 30 minutes, even though the complete session, including instructions, debriefing, practice trials, etc., might run for an hour or more. Even within a 30-minute session, pauses or breaks to help reduce fatigue can be included if deemed necessary from feedback during pilot sessions; however, if the task must be longer than 30 minutes, breaks should be built into the structure of the session. If a study consists of more than one hour of testing, it is advisable to consider breaking it up into multiple sessions, if possible. Researchers should keep in mind, however, that stretching a study across multiple sessions may produce greater risk that participants will change or adopt different strategies across sessions than they would within a single session. In some cases, the decision to have multiple sessions may be dictated by the participants in the targeted population. For example, if the researcher is working with students on a college campus or with individuals within a specific company, it may work quite well to ask them to commit to several sessions. Conversely, if the individuals must come to a location that is removed from their work or home, it may be easier to have them stay for an extended period of time rather than asking them to return for future testing.

Prior to formal data collection, pilot testing is strongly recommended to validate experimental procedures, to help ensure that the participants understand the instructions, and to test any equipment and software that will be used. This should include double-checking any data storage and back-up systems. A small number of participants (e.g., three to five) from the target population is usually sufficient for pilot testing; however, if problems are discovered in
the procedures, additional pilot testing should be seriously considered. Novice researchers may feel that time spent piloting and debugging a procedure are not well spent; however, such testing may not only lead to higher quality data but may actually result in changes that make the data more readily interpretable.

A common practice that should be avoided is using colleagues or graduate student researchers for the pilot study. While such individuals should certainly be asked to provide feedback about research designs or questions on surveys, they should not be used in lieu of a sample from the participant pool. Normally, colleagues or collaborators will have additional experience and information that will allow them to read into questions information that may not actually appear, or they may know how the application is “supposed” to work. Thus, final feedback about the clarity of the procedure or survey questions can only be obtained from a sample of people from the target population, who in most instances will be inexperienced in terms of the sound application in question.

At the beginning of each experimental session, practice trials should be used to ensure that participants are familiar with the test procedures and that they have the opportunity to ask questions so that they understand the task. It is best if practice stimuli are similar, but not identical, to the actual stimuli used in the study. The optimal number of practice trials for a given study can be determined by considering previous research in the area, feedback from pilot testing, and the researcher’s expertise. For some types of study, it may also be important for participants to first listen to the full set of stimuli if they will be asked to perform any type of comparative task (i.e., paired comparisons and sorting tasks). Exposure to the stimulus set assures that participants know the complete reference set of stimuli prior to judging the relations among members of the set. In some cases, such as those involving stimulus sets with relatively unfamiliar or complex information (e.g., auditory data displays), it may even be helpful to present sample auditory displays simultaneously with more familiar equivalent visual analogies (e.g. charts or graphs) to help familiarize the participants with the structure of the auditory displays they will be evaluating.

As a final general recommendation about experimental design, investigators should keep in mind that they are often seeking participants’ subjective perceptions of the stimuli. In most cases, it follows that participants should be instructed to respond as they deem appropriate and that there are no absolutely right or wrong responses. Moreover, every attempt should be made to motivate participants to actively participate in the task, including appropriate remuneration. This may seem counterintuitive to the notion of the “detached” experimenter working within a laboratory setting, but it can have a large impact on the quality of the data procured from perceptual studies.

### 6.2.5 Ethical Treatment and Recruitment of Participants

Investigators who have limited experience with data collection from human participants should make sure that they are knowledgeable about issues relating to ethical treatment of subjects that are mandated by governmental and granting agencies within their countries, as well as human research policies specific to their research settings. In academic and research institutions in the United States there will usually be an institutional review board (IRB) that will have procedures clearly outlined for submitting applications to receive approval
Researchers at other types of institutions or settings that do not normally conduct research with human subjects should check with their institution and seriously consider collaborating with a colleague who has expertise in this area.

One of the most important considerations in designing an auditory display study is for the researcher to select or recruit participants that will be representative of the population that is targeted for use of the type of display being developed. Most of the time, researchers will be interested in the normal adult population with normal hearing. It is interesting to note, however, that very few studies actually include a hearing examination to verify whether the participants have hearing that falls within the normal range. With the increase in hearing deficits that have been documented due to environmental noise (Bauer, Korper, Neuberger, & Raber (1991) and the use of portable music devices (Biassoni et al. 2005; Meyer-Bisch, 1996), researchers should determine whether they need to include hearing testing or whether they may need to restrict the range and types of sounds they use for specific groups. Researchers may also be interested in designing auditory displays for specialized groups, such as children, the elderly, or people with visual impairments. In such cases, it is imperative that the participants reflect the relevant characteristics of the target population (for example, see Oren, Harding & Bonebright, 2008). It can be tempting for researchers to think that they can anticipate the needs of such groups, but this assumption should be quickly questioned. It is best if the research group includes at least one member of the desired target group as a consultant or full collaborator from the beginning of the project, if at all possible, in addition to actively recruiting individuals with the desired characteristics for the most valid test results.

It is also important to consider other general subject characteristics, such as gender, age, and type and level of relevant expertise that might impact the use of the auditory display. For example, there may be differences in the aesthetic value of certain sounds across age groups (see Chapter 7), or an expert user of a specific piece of equipment may be better able to accommodate the addition of sound. It is also important to keep in mind cultural differences that might impact the interpretation of a specific sound (Schueller, Bond, Fucci, Gunderson, & Vaz, 2004) or the perceived pleasantness of sounds (Breger, 1971). Researchers also need to consider that there are other individual differences within populations that may not be so readily apparent on the surface, but which may have dramatic impacts on participants’ abilities to interact with an auditory display. For example, some individuals suffer from amusia, which is a disorder of pitch discrimination and melodic perceptual organization. Such individuals may appear normal in terms of performance on a standard hearing test that is based on simple detection of tones, yet be highly impaired in their ability to recognize melodies or detect changes and harmonic distortions in tone sequences that individuals with normal auditory ability can discriminate with ease (Marin & Perry, 1999). Recent studies (Hyde & Peretz, 2004; Peretz et al., 2002; Peretz & Hyde, 2003) suggest that approximately 4% of the population may have an inherited variety of amusia, while an additional (possibly larger proportion) may suffer from an acquired variety of amusia due to cortical injury related to stroke, trauma, or other pathological conditions (Sarkamo et al., 2009). Designers of auditory displays should thus recognize that just as color deficiency may prevent some individuals from effectively using certain advanced visualization designs, a similar circumstance may exist for the usability of auditory displays by a small proportion of

3The American Psychological Association (www.apa.org) or the National Institutes of Health (www.nih.gov) are good sources for information on ethical treatment of human subjects.
the population who have amusia or related deficits.

Another individual difference that researchers in this area have considered is musical ability, since it seems logical that musical expertise should have an impact on auditory perception, which would consequently affect how individuals interact with auditory displays. Studies that have included this variable have revealed inconsistent results for differences between musicians and non-musicians on basic sound perception tasks (i.e., Beauvois & Meddis, 1997; Neuhoff, Knight, & Wayand, 2002; van Zuijen, Sussman, Winkler, Naatanen, & Tervaniemi, 2005). One reason for these inconsistencies could be that the method for determining musical ability is not standardized. It is also possible that researchers interested in auditory display should be measuring some other construct, such as the basic sensitivity to sound qualities. This issue can only be settled through systematic research examining both basic musical ability and what types of auditory perception (such as auditory streaming, ability to follow simple tonal patterns, sensitivity to rhythm, etc.) are relevant for designing effective auditory displays. There have been some efforts to provide a better measure of auditory perception for elements specific to auditory displays (Edwards, Challis, Hankinson, & Pirie, 2000), but currently such tests have not been widely circulated or accepted within the field.

6.2.6 Sample Size and Power Analysis

Another important research design topic is the number of participants needed, called the sample size. Researchers need to consider the overall task context, which includes the number and type of stimuli, design type, and required statistical analyses to determine the appropriate sample size. These issues can be addressed by reviewing past research in the area to determine the number of participants used or pilot studies can be performed to help make this decision.

When researchers are interested in comparing stimuli, they must choose whether to use a between groups (different participants in each condition) or within groups (the same participants in all conditions) design. Within group designs have the advantage of needing fewer participants and of better statistical power since the variation due to differences in individual participants is statistically removed from the rest of the variance. Therefore, researchers typically make this decision by considering whether there would be any type of carry-over effect from one condition to the other. For example, in a study designed to investigate the effects of mappings for auditory graphs (such as signifying axis crossings with timbre changes versus momentary loudness changes), practice with one mapping would probably affect participants’ ability to learn the other mapping scheme. In such circumstances, a within groups design will not work well, and a between groups design should be used.

Researchers should also consider performing a statistical procedure, called power analysis, which is designed to specify the number of participants needed to get a statistical result that allows for any real effect present to be detected. With insufficient power, results may not be significant simply due to the lack of sufficient sample size rather than that there is no effect to be found. There is also the possibility of having too many participants, which results in trivial effects revealed during statistical analysis, although the most likely error for researchers to make is to have a sample size that is too small rather than too large. The sample size needed for a given study depends on the type of statistical test that will be
used, the number of stimuli to be tested, and the alpha level \(^4\) that will be used to determine significance. There are a number of excellent resources available that present both conceptual background information and practical considerations for performing power analyses (Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990, Lenth, 2001) and there are also a number of software packages, both commercial (i.e. SPSS and SAS) and freeware, that are available for use (Thomas, 1997). More complete information about design type and other related analysis and statistical topics is presented in sections 6.4 and 6.5.

### 6.3 Data Collection Methods for Evaluating Perceptual Qualities and Relationships among Auditory Stimuli

Five commonly used methods for exploring the perceptual qualities of auditory stimuli and their relationships to one another are identification tasks, attribute ratings, discrimination trials, dissimilarity ratings and sorting tasks (see Table 6.1 for a summary). As discussed in previous sections, the researcher should consider each technique in relation to the goals of the project as well as the desired analysis technique to determine which ones are appropriate. This discussion is presented to provide basic information about these techniques; further discussion about how these techniques fit within particular types of analyses will be presented in sections 6.4 and 6.5.

#### 6.3.1 Identification Tasks

Identification tasks for auditory stimuli provide a measure of accuracy for determining whether participants can recognize and label sound stimuli. Normally such tasks provide data that show the percentage of participants who correctly identified the stimulus. Some researchers also collect reaction time data, which is assumed to be a measure of the amount of processing or cognitive effort it takes to complete the task. A short reaction time may indicate that the stimulus represents a well-known and/or quickly identifiable sound, or that a participant made “false starts”. In contrast, a long reaction time may indicate that a sound is unfamiliar or that the participant has lost focus. If a researcher is testing the veracity of synthesized sounds, a long reaction time may indicate that the sound is not a convincing replication of the actual sound. Thus, it is suggested that researchers examine their data to determine whether the pattern of reaction times suggests that outliers are present or whether there is information in these data relevant to the stimulus quality.

Identification tasks for auditory displays include trials that require participants to listen to an auditory stimulus and respond either in a free-form or open-ended format with a written description or by selecting a response from a provided list. In some studies, it is best if the participants are allowed to play the sounds as many times as they desire with no time limit. In such cases, data can also be collected on the number of times the participant played each sound in addition to other measures. If the researcher wishes to obtain intuitive responses, participants are not allowed to change their responses and a time limit may also be imposed.

---

\(^4\)The alpha level is the value set to determine if an obtained result is statistically significant or if it happened by chance. Typically alpha levels are set at .05 or lower, which minimizes the probability of rejecting the null hypothesis when it is true (called Type I error).
<table>
<thead>
<tr>
<th>Task</th>
<th>Typical Measures</th>
<th>Typical Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification</td>
<td>- Accuracy (% correct)</td>
<td>- Design and selection of sounds that are perceptually distinct from each other or that are inherently “meaningful”</td>
</tr>
<tr>
<td></td>
<td>- Reaction time for correct ID</td>
<td></td>
</tr>
<tr>
<td>Attribute Ratings</td>
<td>- Rating scale (e.g. 1 to 7, where 1 = ‘very unpleasant’ to 7 = ‘very pleasant’)</td>
<td>- Discovery of relationships between perceptual and acoustic properties of sounds.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- “Labeling” dimensions that determine similarities and differences between sounds.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Input data for factor analysis and other techniques for determining “structure” among a set of sounds.</td>
</tr>
<tr>
<td>Discrimination</td>
<td>- Accuracy (% correctly compared)</td>
<td>- Design and selection of sounds that are perceptually distinct from each other.</td>
</tr>
<tr>
<td></td>
<td>- Errors (number and type of incorrect responses)</td>
<td></td>
</tr>
<tr>
<td>Dissimilarity Ratings</td>
<td>- Numeric estimate of similarity between pairs of sounds.</td>
<td>- To determine which sounds are highly similar (possibly confusable) or distinct.</td>
</tr>
<tr>
<td></td>
<td>- Dissimilarity or proximity matrix</td>
<td>- Input data for cluster analysis and MDS for determining perceptual “structure” of set of sounds.</td>
</tr>
<tr>
<td>Sorting</td>
<td>- Dissimilarity or proximity matrix</td>
<td>- To determine which sounds are highly similar (possibly confusable) or distinct.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Input data for cluster analysis and MDS for determining perceptual “structure” of set of sounds.</td>
</tr>
</tbody>
</table>

Table 6.1: Summary table for data collection methods.
When the data collected for identification tasks are in an open-ended format, content analysis of the responses is required initially to determine if the meaning of the responses shows a pattern across participants. For example, when Miner (1998) asked subjects to identify a synthesized sound, the responses ‘running river water’, ‘water running in a river’, and ‘river noise’ were aggregated into a single term with a response count of three. The terms ‘rain falling against a window’ and ‘rain’ were not aggregated because the first term provided additional information that would be lost if it were combined with the simpler term ‘rain’. It is important to note that even though this type of linguistic/semantic analysis can be conducted automatically with commercial and non-commercial packages, the researcher will still need to make fine distinctions manually in some cases as noted in the previous example. For both open-ended and fixed format responses, the resulting frequency data can be used to determine whether the participants correctly identified the sounds as well as determining which sounds were confused with each other. Such information can be especially useful for sound designers, since systematically confused sounds can be used as a basis to simplify and speed up the production of synthesized sounds for use in computer software and virtual reality environments (Cook, 2002 and see Chapter 9).

### 6.3.2 Attribute Ratings

Attribute ratings, also called semantic differential ratings, provide information about the perceptually salient qualities of auditory stimuli and are routinely used by investigators working with auditory displays. Researchers using attribute ratings are interested either in understanding the basic perceptual aspects of sound or in combining these data with other analysis techniques, such as factor analysis and multidimensional scaling (MDS), to provide a richer and more complete interpretation of the perceptual structure of the stimuli.

The researcher needs to determine what the appropriate attributes are depending on the type of stimuli used and the purposes of the sound application. Many times these will include a basic set of attributes, such as perceived loudness and pitch, although many other attributes can be used as well, such as roughness, annoyance, or pleasantness. The attributes also need to be clearly and consistently understood by the target population. Finally, the choice of the rating scale varies among researchers but typically semantic differential scales of 5, 7, or 9 points are preferred.

Analysis procedures for rating scale data consist of standard descriptive statistics as well as correlational analysis, analysis of variance, factor analysis, and as an additional measure for interpreting MDS solution spaces. A more complete discussion of these techniques will be presented in sections 6.4 and 6.5.

### 6.3.3 Discrimination Trials

For designing applications that use multiple auditory signals, it is important to determine if people can discriminate between the selected sounds and to measure the extent to which the sounds can be distinguished using a discrimination task. The procedure for a discrimination task requires participants to listen to two sequential stimuli (A and B), which are then followed by a third stimulus (X). Participants are then asked to determine if X is the same as A, B or neither of them (Ballas, 1993; Turnage, Bonebright, Buhman, & Flowers, 1996).
In the instructions, participants are informed that there will be a number of ‘catch’ trials on which the correct response would be neither. These trials are necessary to make sure that participants are attending to both stimuli A and B before making their judgments rather than adopting the simpler strategy of ignoring A, attending to B, and making a same-different judgment for the B-X pair (Garbin, 1988).

Basic analyses of data from this procedure consist of comparisons of correct responses and errors using descriptive statistics, such as means, standard deviations and ranges. Investigation of the types of errors for individual participants and for composite data from all participants can be examined for patterns that indicate perceptual similarity among the stimuli.

6.3.4 Dissimilarity Ratings

Dissimilarity rating 5, also referred to as proximity rating, paired comparison, or similarity judgment, is when participants provide a numerical assessment of dissimilarity for each possible pair of stimuli in the set. A typical dissimilarity rating task might instruct participants to “listen to each pair of sounds presented and to rate their degree of dissimilarity by using a 7-point rating scale where 1 = extremely similar and 7 = extremely dissimilar.” It is also possible to have participants make a mark along a continuous line with labeled endpoints to indicate degree of similarity. A number of commercial data collection or survey software packages can be used for participants to enter such judgments. However, paper and pencil forms may also be used for participants to enter the ratings.

Regardless of how the rating data are recorded, considerable attention should be given to the manner in which the sound samples are presented. Comparing pairs of sounds presents some cognitive and perceptual issues that differ from those encountered with comparing visual displays. As discussed previously in this chapter, sounds must be listened to sequentially, which means that there is a memory component to the comparison task that would not be the case for simultaneously displayed visual stimuli. There may also be order effects, such that a similarity rating may be slightly different for the same pair, depending on which sound is played first. There are several choices for how one might present sound pairs to address these potential complications. One method of dealing with order effects is to present each pair twice - once in each order. The mean of the two ratings can be used as a participant’s estimate of similarity between the pair of sound samples. If this procedure is followed, however, there will be a minimum of $N \cdot (N - 1)$ ratings performed by each participant, where $N$ is the number of sounds in the set. This procedure may produce a time consuming (and perhaps arduous) task if the number of stimuli is large. For example obtaining dissimilarity ratings for 50 different automobile horn samples would require presentation of 2450 pairs of horn toots to each participant. An alternative approach to dealing with potential order effects would involve randomizing the order of each pair and the order in which each pair is presented during the session for each participant, thereby cutting the number of pair presentations in half: $N \cdot (N - 1)/2$ comparisons. A third alternative approach to addressing pair order effects involves allowing the participants to listen to each member of a pair in any order as many times as they wish before entering a dissimilarity rating. This can be achieved by presenting, on each trial, a pair of clickable icons that elicit each of the sound samples.

---

5For both dissimilarity ratings and sorting tasks, researchers should be mindful of differences among the stimuli, such as duration or amplitude that could perceptually overwhelm other attributes of the stimuli the researcher may wish to examine. In such cases, the stimuli could be equalized on the relevant dimensions.
along with an icon or button that presents a dissimilarity rating box that can be activated when a participant is ready to respond. Since participants may choose to listen to each sound more than once under this procedure, it may take slightly longer to complete than the fixed schedule with randomized order of pairs, and it does not give the researcher full control over the number of times participants are exposed to the stimuli. Investigators who plan to collect dissimilarity ratings among sound samples must thus weigh the costs and benefits of these alternative approaches for their particular research or development project (and perhaps also consider the option of assessing perceptual dissimilarity by using a sorting task, which will be described in the next section).

The data obtained from a dissimilarity rating task are typically configured into a “dissimilarity” matrix for each participant in which cell entries (assuming a 7-point rating scale) will vary between 1 and 7. Most computer programs for clustering or scaling such data require (or at least accept) a “lower left triangular” matrix (examples and more information about these techniques will be shown in section 6.5) for the input of such data, often called “proximity” data.

6.3.5 Sorting Tasks

An alternative method for obtaining perceptual distance or dissimilarity ratings among stimuli is a task in which participants sort a set of stimuli (typically 20-80 examples) into “piles” or “groups.” Traditionally, such methods have been used for visual and tactile stimuli (Schiffman, Reynolds, & Young, 1981); however studies indicate their utility in investigating auditory stimuli as well (Bonebright, 1996, 1997; Flowers et al., 2001; Flowers & Grafel, 2002).

While sorting is not an activity normally associated with sounds, current technology makes it quite easy to collect sorting data on sound samples by presenting participants with a computer screen folder containing numbered or labeled icons that activate the presentation of a sound file. Participants are allowed to click on each icon as often as they wish to listen to it and to move the icons into different locations on the screen based upon their judgments of similarity until they are satisfied that they have formed meaningful groupings. The experimenter then records the group each stimulus was placed in (a process which could be automated by software that senses the screen position of the final icon locations). A dissimilarity matrix is generated for each participant by assigning the value “0” to each pair of stimuli that are sorted into the same pile, and the value of “1” to each stimulus pair that is sorted into a different pile. Logically, this is equivalent to obtaining dissimilarity ratings using a “two-point” rating scale for each participant, as opposed to the typical seven point scales used in dissimilarity rating tasks. As with actual dissimilarity ratings, one may sum these matrices across participants to obtain a group or “composite” dissimilarity matrix. Each cell entry of this composite matrix thus consists of an integer that is the count of how many participants assigned a particular stimulus pair to different piles. The composite matrix may be submitted for clustering or MDS procedures, and individual participant dissimilarity matrices may be reconfigured into linear vectors and submitted to reliability analysis programs. The authors of this chapter have developed simple software routines to perform this transformation. One version writes the lower triangular dissimilarity matrix (and sums these matrices across participants to provide a composite or group dissimilarity matrix), while the other version “stretches out” the individual participant dissimilarity ratings for each pair into a linear vector so that similarity of sorting patterns among subjects can be assessed by correlation and
The following example (shown in Figure 6.1) illustrates the process of transforming sorting data into lower triangular dissimilarity matrices. Suppose there are three participants who each sort ten different sound samples (note that in an actual design scenario, there would likely be far more participants and stimuli). The three rows of ten digits on the left of Figure 6.1 represent the sorting data from each participant. The cell entries are the “pile number” in which that particular subject sorted each stimulus. For example, the first subject assigned stimuli #1 & #4 to pile 3, #2 and #3 to pile 1, #5, #6, and #10 to pile 4, and #7, #8 and #9 to pile 2. The “pile numbers” are arbitrary and need not correspond across participants, since the matrix and vector data only reflect whether each pair of stimuli were grouped together or not. Note that while the first two participants used four piles, the third participant only used three piles. For the first participant, pairs 1-4, 2-3, 5-6, 5-10, 6-10, 7-8, 7-9, and 8-9 should all receive a zero in the lower left triangular dissimilarity matrix (where the first number in the pair is the column and the second the row), and the remaining cell entries should be “ones”. The right side of Figure 6.1 displays the three lower triangular dissimilarity matrices computed for these three participants, followed by the group (summed) matrix at the bottom. Note that in these matrices there is also the matrix diagonal (each entry a zero) since this is required by several popular data analysis packages, such as SPSS, when submitting dissimilarity data to perform multidimensional scaling or cluster analyses.

```
3 1 1 3 4 4 2 2 2 4
3 2 2 1 1 1 4 4 4 1
1 1 1 2 3 3 2 1 2 3
```

Figure 6.1: Transforming sorting data into individual and composite dissimilarity matrices.

Unless the participants hear all the stimuli first, it is probably best to allow at least two sorting trials for a set of stimuli, and to use the matrices generated by the final sorting for further reliability analyses.
analysis. The completion of the first sorting trial is necessary to familiarize each participant with the presence and range of variation among auditory attributes and features within the set of stimuli.

It should be noted, that when collecting sorting data, researchers sometimes place weak constraints on the number of groups (e.g., “no fewer than three, no more than eight”), and/or a restriction on the minimum number of stimuli (e.g., “2”) that can be included in any group. In the opinion of the authors (and particularly for sonification design purposes), specific instructions of this type are of little value. If one stimulus is truly unique, participants should be allowed to indicate that. Additionally, failure to separate a relatively diverse set of ten or more stimuli into at least three categories rarely happens.

One clear advantage of sorting tasks over dissimilarity rating tasks is the speed with which the data can be obtained. It is much quicker and much less tedious for participants to sort stimuli into piles that “go together” than to be presented with at least \( N \cdot (N - 1)/2 \) numeric rating trials. Another advantage sorting tasks have in relation to dissimilarity rating tasks is that once the participants have finished sorting the stimuli into groups, they can be asked to label each of the categories. Such information can help the researcher understand what the participants explicitly thought they were doing, and it may also help in interpreting the results of the data analysis. However, it should be noted that participants may be using strategies of which they are not consciously aware, which the data analysis may be able to expose. When used in conjunction with techniques that correlate physical stimulus properties with positions in a multidimensional scaling plot or a cluster plot (examples of which will be discussed in section 6.5.2, below), labeling data may be quite instructive.

### 6.4 Analysis of Data Obtained from Identification, Attribute Rating, Discrimination, and Dissimilarity Rating Tasks

There are two categories of data analysis approaches, correlation based, and group or condition comparison based, which are often helpful in making decisions about the effectiveness of display properties. Correlation based analyses are performed when one wishes to determine the strength of relationship between two (usually continuous) quantitative variables (for example between pitch of a data stream in an auditory graph designed to display temperature, and observers’ estimates of temperature). Group (or condition) comparison analyses are used when one wishes to see whether two or more different conditions produced different values (usually based on the mean) of some quantitative measure. For example, if one has two alternative designs for an alarm earcon in an industrial display, does one produce faster response times than the other? It is often the case that a designer will find it useful to employ both correlation based and comparison based analysis procedures during the process of designing an application involving auditory displays.

#### 6.4.1 Correlation Analyses

While there are several different statistical measures of correlation, one of the most commonly used is the Pearson’s correlation coefficient, \( r \). In this analysis, two quantitative variables,
labelled $X$ and $Y$, are tested to determine whether there is a linear relationship between them. The correlation coefficient $r$, provides information about the direction (whether $X$ and $Y$ vary in the same direction or vary in opposite directions, indicated by whether the computed value of $r$ has a positive or negative value) and the strength of the relationship (a value of $+1.0$ or $-1.0$ indicate “perfect” linear relationships, and 0.0 indicates no relationship). For most designers of auditory displays, correlation analyses, by themselves, will not provide sufficient information for product evaluation without reliance on additional approaches such as group or condition comparisons. However, Pearson’s correlation coefficient lies at the heart of more sophisticated analyses (such as regression analysis, factor analysis, cluster analysis, and MDS) that can be used effectively to determine the perceptual and acoustic qualities among sets of auditory stimuli. (For detailed statistical discussion of the Pearson’s correlation coefficient and multiple regression readers should consult Cohen (2003) or Pedhazur (1997); for factor analysis, Gorsuch, 1983 or Brown, 2006 and for questions about multivariate analyses, Tabachnick & Fidell, 2006).

6.4.2 Comparing Conditions or Groups: t-tests, ANOVA and Related Procedures

Most researchers designing auditory displays will wish to compare users’ responses to individual sounds or to different display formats to determine which one would work best for a particular application. Such studies may use measures of performance, such as users’ speed or accuracy in responding to an auditory display (see Bonebright & Nees, 2008 for an example study that uses both types of measures), or they may use subjective attribute ratings, or similarity ratings (for example to determine whether one set of auditory icons “matches” a set of visual icons better than another set of icons). The data from these studies is then submitted to an analysis technique that compares the means of performance measures or ratings among the conditions.

Two basic statistical procedures for evaluating differences between means are Analysis of Variance (ANOVA) and the t-test. Each of these has versions (with different computational techniques) suitable for comparing means among conditions in between-group and within-group designs (see section 6.2.6). The present discussion will focus on ANOVA since this technique is much more flexible and can test differences among multiple subject groups across multiple conditions while the t-test can only be used for testing differences between means from two conditions. Software routines for performing analyses using ANOVA and related techniques are available in a wide range of data analysis software packages (e.g., SAS, SPSS, R) and some limited capabilities for using these techniques are embedded among the “add on tools” in “professional” versions of popular spreadsheet programs such as Microsoft Excel. While a full discussion of ANOVA (and more advanced related techniques such as

---

6Researchers should keep in mind that there are other types of non-linear relationships between quantitative variables (such as quadratic), which will not result in a significant correlation coefficient. It is always valuable to plot the data points for a visual representation of the data to understand more fully the relationship. Please consult the references specified in the text for information on this and other issues related to the pitfalls of correlation analysis.

7There are a number of good references for multivariate statistics. The current recommendation for Tabachnick and Fidell is based on the applied nature and readability of their book. However, there are other references, such as Johnson, 2002; Johnson & Wichern, 1998; Hair, Tatham, Anderson, & Black, 1998, Abdi, Edelman, Valentin, & Dowling, 2009, that also provide excellent coverage of the topic.
MANOVA and ANCOVA) is beyond the scope of this chapter (interested readers should consult Tabachnick and Fidell, 2006 or one of the other references presented in footnote 7), a general discussion of the basic logic of statistically evaluating differences among means, as well as some of the common pitfalls encountered in applying and interpreting such analyses are in order.

The logic behind ANOVA and related statistical procedures is to compare the variance among the values (e.g., performance scores, ratings) within each group (condition), considered the error variance, with the variance between the groups or conditions (which is presumed to reflect both group differences and error variance). This comparison is done by computing a ratio, called an F-ratio. If the resulting F value differs more than would be “expected by chance”, the “null hypothesis” that the means of the conditions or groups are the same is rejected, and the researcher then has some evidence that there are actual differences between the groups on the variable of interest. For a study that has more than two groups or conditions, follow-up analyses are needed to determine exactly where the differences lie since the “omnibus” F test only determines that there is a difference among the groups, but not which group differences are significant. Clearly, the researcher must also examine the values of the group or condition means to determine the direction of the differences, should the F value be significant. It is important to note that there are conflicting opinions about post hoc comparisons and their appropriate use. Researchers should check the previously mentioned references on multivariate statistics as well as consult with colleagues in their respective disciplines to determine which method is the best to use for publication purposes. It is also important to point out that display design and optimization decisions are not the same thing as pure scientific research being prepared for a research journal; thus, ultraconservative statistical tests may not be necessary.

6.4.3 Caveats When Using Techniques that Compare Groups

Even though comparing means by ANOVA and related techniques is relatively simple to perform with statistical software, there are pitfalls that should be recognized and avoided, particularly by researchers who have little or no prior experience with applying these techniques. One typical problem results from failure to screen data prior to performing mean comparisons. Not only are there common problems with data sets, such as missing data points or data that contain errors due to data entry or to equipment or software problems, but there are also more subtle issues that should be examined. Averaged data and particularly computations of variance are extremely sensitive to outliers, and if a data set contains them, the means can be either artificially inflated or deflated leading to finding a difference that doesn’t really exist for the population of interest or not finding one that is there. There is also the possibility that the sample of participants may be made up of multiple populations, such as people who process sounds differently than others, or it could be that the outliers are participants who misunderstood the instructions. In the first case, it would be good for the researcher to be able to identify this sub-group so that appropriate accommodations can be made for them when they use the sound application. In the second case, typically indicated by a number of outliers who share no common pattern of responses, it is extremely difficult to determine if these occur due to general perceptual difficulties encountered with the displays, or to basic inattention to the task. Interviews or post testing surveys of individual participants may provide some guidance in this regard. In general, the presence of substantial lack of
reliability among participants in responding to auditory displays should trigger a note of concern about whether a design has been adequately optimized. In the case where there is additional, clear empirical evidence that some participants did have difficulty understanding the task or were inattentive, the data can be removed from the data set before further analyses are performed and a full explanation for this action included in any manuscript written for publication purposes.

It is most important to note that neither data entry errors, nor presence of outlying data observations are likely to be easily discovered without an initial data screening. Screening can consist of simple visual inspection of data values in a table or spreadsheet if the number of data records is relatively small, but for larger data sets some type of software assisted screening should be considered. In some cases reliability analysis routines may be useful (an example will presented later in section 6.5.1), and it is possible that some types of visualization schemes, such as plotting condition profiles for each participant on a common plot to see if any visually “jump out”, may also be helpful (Wegman, 2003). It should also be noted that even sonification of raw data values by mapping them to pitch (perhaps organized as profiles of observations from each participant) could be useful in pointing out anomalies in the data prior to formal statistical analyses.

Another pitfall researchers should be wary of is the difference between a statistically significant difference and a practical difference. If the analysis finds that the difference between the group means was significant, the researcher can assume that the difference most probably didn’t happen by chance. But the probability value (typically set at less than .05) doesn’t state what the actual effect size is. In order to determine this, additional statistical tests, such as $\eta^2$ or $\omega^2$, which provide an estimate of the proportion of variance due to the differences in the conditions, need to be performed (Tabachnick & Fidell, 2006). However, even if there is a significant difference and the effect size is large, the difference between the means may not be practically large enough to matter when the sound application is used in a real world setting.

The final pitfall that researchers should keep in mind occurs when there are multiple comparisons being performed within a given study. Alpha inflation or Familywise type I error (FWER) occurs when each comparison performed has the probability of .05 that the null hypothesis was rejected when it should have been retained. For each additional analysis, the probability of committing this type of error increases by the amount of the probability value used. The issue of adjusting for alpha inflation is controversial, and there are a number of methods (such as Scheffe, Tukey, Dunnett, Bonferoni or Fisher tests) that can be used ranging in how conservative they are, that will correct for the type I error rate (Keppel & Wickens, 2004). Obviously, these corrections decrease the likelihood of finding a significant difference; however this is justified since the convention is to be conservative in terms of stating that differences exist.\footnote{There is a movement in a number of disciplines to use statistical techniques, such as Bayesian statistics, that do not have the disadvantages of null hypothesis significance testing (for a discussion of this issue, see Kruschke, 2010 or Wagenmakers, Lodewyckx, Kuriyal and Grasman, 2010). However, statistical testing as described in this chapter is still the predominantly accepted method.}
6.5 Using “Distance” Data Obtained by Dissimilarity Ratings, Sorting, and Other Tasks

Evaluation of the overall usability of an auditory display requires consideration of both the effectiveness of the perceptual mappings between sound and information that the designer intends to present, and the reliability of perception of the display among potential users. Perceptual mappings play a critical role in making sure that the listeners extract the desired information from the display. For example, if the designer wishes to present data values that are increasing, pitches that increase would be appropriate. However, if the designer also adds changes in loudness to this auditory stream, the interaction between changes in pitch and loudness may lead to “distorted” estimates of the magnitudes since changes in pitch can affect judgment of loudness and vice versa (see Neuhoff, Kramer, & Wayand, 2002). Such a display could be described as reliably perceived, since all the participants may perceive the graph in exactly the same way, but its ability to display the underlying information would be compromised. Alternatively, an auditory graph of data that appears to faithfully represent the structure of data to about 40% of users, but conveys little or no information to the remaining 60% (or, worse yet, conveys a totally different structure among a subset of users), would have serious reliability shortcomings, and thus its overall usability would also be low.

The use of data collection techniques that generate “perceived distance estimates” among auditory display elements can be used to address the issue of consistency of perception among users via reliability analysis, and produce descriptions of the actual perceptual relationships among the display elements via techniques such as cluster analysis and MDS. Solutions from clustering or MDS routines may then be examined to determine whether they meet the objectives for which the display is being designed. For example, if display elements are auditory graphs representing multivariate data, one can make statistical comparisons between values of variables in graphs included in different clusters, and/or one can use regression analysis to determine the relationship between numeric values of variables and the position of the graphs in an MDS structure (e.g., Flowers & Hauer, 1995). If the display elements are real or synthesized “product sounds”, one can use such procedures to determine relationships between acoustical properties of sounds and user perceptions to guide design or predict consumer preferences.

There are several methods commonly used to assess the perceived “distance” or dissimilarity between stimuli for purposes of clustering or scaling. “Direct” methods include the use of dissimilarity ratings and sorting tasks, which were discussed in sections 6.3.4 and 6.3.5. Perceptual dissimilarity between stimuli can also be measured “indirectly” by computing it from attribute rating tasks, which were discussed in section 6.3.2. However perceptual dissimilarity measures can also be computed from measures of performance (speed or accuracy) from tasks requiring participants to make perceptual discriminations between different stimuli, such as same/different judgments of stimulus pairs, or speeded classification (e.g., “press the right key if you hear sound A; press the left key if you hear sound B”). The “direct” methods (dissimilarity rating and sorting) offer a considerable advantage in the speed of data collection and are probably preferable for most applications involving evaluation of auditory displays.
6.5.1 Using Reliability Analysis to Assess Dissimilarity Rating or Sorting Consistency among Participants

Reliability analysis is a technique typically used for the assessment of test items used in educational or psychometric tests. Presented here is the use of reliability analysis using sorting or dissimilarity rating data for obtaining a measure of “agreement” among the participants about the perceptual structure of a set of sounds that might be used in an auditory display. For this purpose, the “test items” are each participant’s “stretched out dissimilarity matrix”. There will be \( N \cdot (N - 1)/2 \) of entries in each vector, where \( N \) is the number of stimuli that were sorted. Examples of such vectors could be generated by taking each of the individual matrices on the right side of Figure 6.1, eliminating the zeros that make up the diagonal, and then lining up the data in a separate column for each participant.

Reliability analysis routines, such as SPSS Reliabilities, compute several measures of reliability and scaling statistics, but for the present purposes, an overall measure of consistency among participants, the Cronbach’s alpha (Cronbach, 1951) would be used. It would also be necessary to have a measure of consistency (correlation) between each participant’s vector and the composite of the entire group of participants for detecting participants whose sorting patterns substantially depart from the overall group sorting pattern (outlier detection). For example, SPSS\(^9\) provides a printout column titled “Item-Total Correlation” that presents this information for each participant (the participants are the “items” in the application), as well as a column titled “Alpha if Item Deleted” (a listing of what the overall reliability would be if a participant or “item” were to be excluded from the analysis). Alphas above 0.70 indicate quite reasonable agreement or consistency, with values above 0.80 providing a quite high level of confidence that there is a solid shared basis of judgments of similarity among the stimuli.

It is reasonable to expect on the basis of distribution of pitch discrimination impairment and other auditory deficiencies in the general population that a small percentage of participants will have difficulty with discriminating pitch changes normally encountered in music (Marin & Perry, 1999) or may simply fail to understand the nature of the sorting task. In practice, it is difficult and not overly useful to distinguish between these two types of participants. One can adopt a policy of excluding participants whose grouping patterns exhibit a negative correlation or a correlation of less than some small positive value (e.g., +0.05) with the remainder of the group (as indicated by the Item-total correlation) from inclusion in subsequent MDS or clustering analyses, on the basis that they are outliers and are not likely to be representative of the population for which the auditory displays will be used, particularly if the alpha after exclusion is substantial in size.

Reliability analysis of sorting patterns can also be useful as a general performance measure for making comparisons among different display designs or durations for purposes of optimizing display design. For example, Flowers & Grafel (2002) had participants sort two sets of auditory graphs representing monthly samples of climate data. One set, the “slow” displays, presented 31 days of weather observations in 14.4 seconds, while other (“fast”) displays presented the same data in 7.2 seconds. Sorting reliability for the “slow” displays was substantially lower than for the “fast” displays (0.43 vs. 0.65), even though participants indicated a preference for the slow displays, and stated that the fast ones were “too fast to

\(^9\)SPSS is only one of the commercial packages that can be used for the analysis specified here. Such an analysis can also be performed using freeware, such as R or Scilab.
perceive detail.” Subsequent display evaluation, using displays of similar design but with an intermediate duration of 10.0 seconds produced higher sorting reliabilities ranging from 0.71 to 0.84. For these types of auditory time series graphs (which will be discussed in more detail in the next section), display duration clearly affected the consistency of sorting among users.

6.5.2 Inferring “Perceptual Structure” using “Distance” Data: Clustering and Multidimensional Scaling (MDS)

Once it has been ascertained through reliability analysis that participants agree, to a reasonable extent, about which sound samples are similar or dissimilar to each other, the application of clustering and/or MDS procedures can be used to generate displays of perceptual structure among the set of sounds being investigated. A full treatment of either cluster analysis or MDS procedures is beyond the scope of this chapter; readers interested in applying these procedures should consult one or more of the authoritative sources in these areas such as Borg & Groenen, 2005; Davison, 1992; Kruskal, 1977; Kruskal & Wish, 1978; Schiffman, et al., 1981; or Young & Hamer, 1987. However the following discussion should provide some basic guidelines about how these procedures can be used by investigators and designers of auditory displays. Both hierarchical clustering and MDS are data structure display techniques that analyze “distance” data, and provide a display that illustrates perceptual “distance” relationships among stimuli. Both techniques can be used in conjunction with either rating data or acoustical properties of the stimuli to show how these perceptual distance relationships relate to psychological (perceived) or physical stimulus attributes (see Davison, 1992).

To illustrate use of these techniques, examples of data analyses from a previously unpublished study of auditory “weather graph perception” conducted as a follow-up to the study of Flowers and Grafel (2002) are presented here. These data were generated by 30 participants who each sorted a set of 23 auditory graphs into perceptually similar groups. The auditory graphs displayed monthly samples of historical weather observations from Lincoln, Nebraska, obtained from the High Plains Regional Climate Center (www.hprcc.org). These 23 monthly records were selected to cover a representative range of variation in temperature and precipitation patterns typical of the Great Plains of the United States during warm season months across the historical period of 1934-2000 - a period during which substantial climate variation occurred. The auditory displays presented each day’s high and low temperature as an alternating four note synthetic string MIDI stream for which pitch was mapped to temperature. On days in which precipitation occurred, a one to three note MIDI grand piano was imposed over the last half of the four note string sequences to indicate rainfall amount. (For additional details about the display format see Flowers, Whitwer, Grafel, & Kotan, 2001 and Flowers & Grafel, 2002).

The basic display output of a hierarchical clustering procedure is a “tree” structure (sometimes shown as an “icicle plot” or a “dendrogram” depending on one’s display preferences). These displays depict clusters of stimuli that “belong together” under a hierarchical “agglomeration schedule” that adds stimuli to clusters and clusters to each other based on analysis of distance data. There are several choices among clustering algorithms used for determining the criteria for combining groups, and at what “level” the clusters or stimuli are combined. However the objectives of these algorithms are quite similar; in many cases the results they produce are also highly similar. Figure 6.2 displays a dendrogram created by SPSS using the weather
sample sorting data and the average linkage method, which is a typical default clustering algorithm.\(^{10}\)

Determining the agglomeration level at which joined items should be considered “meaningful” to be treated as a group is a relatively subjective judgment. If the clusters that result after that judgment differ meaningfully in properties of the stimuli themselves, or correspond to additional rating or other performance data obtained with the stimuli, one gains confidence that the groupings reflect meaningful perceptual decisions on the part of the participants, and thus they can guide a variety of subsequent design decisions. Figure 6.2 was selected as an example since it happens to illustrate some “extremes” of what might happen (and did happen in this case) and to point to some interesting and informative data features. The visual overview of the agglomeration structure suggests three major groupings at about level 15, but with breaks among these groupings at levels 10 through 13 suggesting that a 5 “cluster” structure would be a reasonable description. However, there is one feature that stands out. One stimulus, the auditory display of the weather from October 1999 (sound example S6.1) does not combine with any group until level 15. This pattern of extreme late combination is suggestive of an outlier - a stimulus that does not belong to any group. Inspection of the weather properties for this month suggest that it was indeed meteorologically unique within the set of 23 monthly climate samples (additional sound examples are referenced directly prior to Figure 6.3). It was exceptionally dry (a trace of rain on each of three days), but quite cool. Coolness and dryness happen to be features that do not conjoin in the other stimuli in this set. Musically, the sonification of October 1999 consisted of an atypical low pitched

\(^{10}\)For more details about different types of clustering analyses, see Johnson, 1967.
temperature stream with only five single high piano plinks representing rain. The two months with which it was combined, at the last resort, were August 1947 (sound example S6.2) and the dust bowl month of July 1936 (sound example S6.3). These were also months of exceptional drought and only three days of rain. But these two months also had searing heat (up to the all-time record 115 degrees Fahrenheit for the region) and would have produced a temperature stream averaging more than an octave higher in pitch throughout the 10-second display. Within the remaining 20 monthly weather samples there were both hot and cool months with either moderate or high amounts of precipitation, but no other cool and very dry months. So it “makes sense” that drought was probably the common attribute that determined October 1999’s final admission to a cluster.

When clusters have been defined by a clustering routine, one may then inspect whether the clusters differ in terms of specific measurable properties of the stimuli or in terms of additional ratings of the stimuli obtained. Provided there are enough members of individual clusters to provide sufficient statistical power, traditional techniques such as ANOVA can be used for that purpose. In the present case, clusters differed significantly in terms of both total precipitation, and number of days on which precipitation occurred. When October 1999 was included in the analysis by clusters, there were no significant differences between clusters in temperature. However, with the removal of the October 1999 an overall significant effect of temperature was found that distinguished among the clusters as well as a pattern suggesting that participants were able to perceive the key meteorological properties of these different historical weather records by listening to them.

The objective of MDS procedures is to provide a spatial depiction of stimulus similarity relationships - typically in Euclidean space. MDS procedures use iterative algorithms to discover a spatial configuration of the stimuli that is compatible with at least the ordinal relationships among the dissimilarity measures among the stimuli - and to do so in a minimum number of Euclidean dimensions. How “compatible” a fit in a given number of dimensions (typically 2, 3, or sometimes 4 for perceptual stimuli) happens to be is usually assessed by at least one, and typically two measures of the “degree of fit” that has been achieved once the iterative routine has determined that it has “done its job”. MDS computation routines such as ALSCAL11 (Young & Lewyckyj, 1979) provide STRESS and $R^2$ as indices of discrepancy between distances among “optimally scaled” points and the positions produced by the final configuration (for a discussion of computational details see Kruskal & Wish, 1978). STRESS ranges between zero and one and is sometimes referred to as a measure of “badness of fit” since poor fits are associated with larger numbers. $R^2$ is a form of a multiple correlation coefficient – in this case between optimally scaled dissimilarities and the MDS model distances. It gets larger as the “fit” of the model improves, and it can be viewed, like other types of multiple correlations, as a “proportion of the variance” of the optimally scaled data that can be accounted for by the MDS solution. Good fit does not imply a meaningful solution however. The user should attempt to achieve a solution in the minimum number of dimensions that produces an acceptable level of fit, since using a large number of dimensions, may lead to small STRESS and large $R^2$ values, but a meaningless “degenerate” solution.

To illustrate an example of MDS applied to assessing perceptual structure of auditory display stimuli, the same example of weather data sonification used to illustrate clustering methods

---

11There are other MDS algorithms [for example, CLASCAL, see Winsberg & De Soete (1993) INDSCAL or MULTISCALE, see Young, 1984)] that can be used, although the discussion of their relevant advantages and disadvantages is beyond the scope of this chapter. See the MDS references for more information.
will be used. The SPSS ALSCAL routine was applied to the dissimilarity data from the sorting task obtained for monthly weather records that generated the cluster display previously shown in Figure 6.2. A “satisfactory” fit was obtained in two dimensions, stress = 0.135 and $R^2 = 0.923$. Figure 6.3 displays the spatial configuration in two dimensions, upon which rectangles encompassing the cluster groupings described in the earlier discussion of clustering are superimposed. Inspection of this display shows that October 1999 (sound example S6.1) again stands out as an outlier - farther apart from its neighbors in the cluster than any other month in the sample. However, the geometric relationships show that the other two members of that ill-defined cluster are spatially close to other brutally hot and almost as dry months such as August 1947 (sound example S6.2) and July 1936 (sound example S6.3). Notably cooler and very wet months, such as June 1947 (sound example S6.4), and May 1996 (sound example S6.5) are on the opposite (left) side of the display. Notably cooler and very wet months, such as June 1947 (sound example S6.4), and May 1996 (sound example S6.5) are on the opposite (left) side of the display (Please refer to four additional sound files for more examples from the clusters in the MDS solution space - August 1960 (sound example S6.6), August 1940 (sound example S6.7), August 2000 (sound example S6.8), and July 1934 (sound example S6.9) as well as nine examples from fall and winter months that were not included in this study – December 2000 (sound example S6.10), December 2001 (sound example S6.11), December 1999 (sound example S6.12), February 1974 (sound example S6.13), January 2001 (sound example S6.14), January 1940 (sound example S6.15), January 1974 (sound example S6.16), November 1940 (sound example S6.17), and November 1985 (sound example S6.18).

Figure 6.3: MDS configuration obtained from auditory weather graph sorting data.

It was previously mentioned that the mean temperature values, total precipitation, and number of days on which precipitation occurred differed significantly among the clusters (at least when the outlier October 1999 was excluded). Figure 6.3 clearly shows that the stimuli defined by these clusters appear in different spatial regions. With an MDS configuration,
one can use multiple regression techniques to indicate the relationship between the positions of stimuli in the space defined by the MDS dimensions and some measured quantitative property of the stimuli, by using the MDS dimensions as predictors and the property value as the dependent measure. In the present case, regression can help identify regions of “wetness versus dryness”, “many rainy days” versus “few rainy days”, and “warm” versus “cool” through predicting total monthly precipitation, number of days on which rain fell, and mean temperature of the month, using the two MDS dimension scale values of each stimulus. The ratio of the beta weights of the two predictors defines the slope of the “best fit vector” for each of these predictors; thus one can draw a line, passing through the origin and use this computed slope to illustrate these relationships. Figure 6.4 displays such vectors. One is only justified in displaying property vectors in this manner if the result of the multiple regression analysis shows that MDS axes significantly predict the stimulus property being represented; in this case all three regressions were significantly predicted. The RSQ values listed on Figure 6.4 are the squared multiple correlation, or the proportion of variance accounted by the regression models. In this particular situation, one can infer that the experimental participants who listened to these auditory depictions of month-long samples of weather observations were indeed sensitive to the sonic representation of temperature and precipitation patterns.

In summary, the combination of clustering procedures with MDS and regression analyses based on stimulus attributes can provide a very useful set of exploratory and visualization tools for discovering perceptual relationships among auditory stimuli, and thereby guide choice or design of auditory display components for a wide range of applications, such as sonified data displays, auditory icons, earcons, status indicators, alarms, etc. These tools can guide discovery of which sounds are perceptually distinct from each other, and which have sufficient similarity that confusability might become an issue. The addition of regression procedures to MDS can help determine the relationships between subjective quality ratings of

![MDS configuration of auditory weather graph data with stimulus attribute vectors included.](image)
sounds, their acoustical properties, and their perceptual similarity structure (see Gygi, Kidd, & Watson, 2007 for another MDS analysis example used in the investigation of environmental sound perception).

6.6 Usability Testing Issues and Active Use Experimental Procedures

When considering auditory displays from a usability perspective, there are a number of issues that the designer and researcher must take into account. These issues are basic to all usability testing and include the time it takes the user to learn the application; the speed at which the user can accomplish tasks with the application; the number of errors that occur while using the application; the ease of retention over time for how to use the application; and the overall subjective satisfaction of the user with the application (Schneiderman, 1998). These issues all point to the value of providing testing that promotes actual use of the auditory display by the target user population in the target environment early and often during the process. Conversely, it is also important that experts play a role so that they can use their knowledge to help winnow down options. For example computer scientists might evaluate the cost effectiveness of specific auditory displays in terms of computer processing power, while perceptual psychologists might consider the cognitive and perceptual abilities of human users in relation to the proposed display.

6.6.1 Testing in the Laboratory versus Testing in Real World Settings

It is also absolutely essential to consider the use of data collected in a laboratory setting in comparison with use in the target environment. Results of a strictly controlled experiment may suggest to a developer that a particular aspect of an auditory display “works well” because it produces statistically significant effects on performance that are in the desired direction. However, this does not indicate that the application will work well in a less controlled environment (e.g., a workplace that has noise that overpowers the sound display or a working environment that results in such sound displays adversely affecting co-workers); thus practical significance needs to be thoroughly examined. In addition, while participants in an experiment may be willing to endure multiple sessions for training purposes due to any compensation they might receive, users in a real environment must immediately see that the potential benefits outweigh any costs in learning or using the display, otherwise they may choose to simply disable the application.

Assessment of sound applications using active-use procedures emphasizes the actual use of the product or application in the “real-world” environment. Such techniques, including surveys, verbal protocols, focus groups and expert appraisals can be used both in the target environment or in a usability laboratory that is set up to provide a comparable environment to the one where the application will actually be used (Jordan, 2002; Nielsen, 1993). In these laboratories, participants can work with the display and provide feedback to the researchers. In this type of testing, it is imperative that the subjects realize that they are not being tested, but rather that it is the application or product that is under investigation.

There are many useful references for usability testing. Schniederman’s 1998 book is a good general reference, but the reader may also wish to consult Dumas and Redish (1993) or Nielsen (1993).
6.6.2 Surveys

Surveys can be designed to collect data before, during, and/or after the participant has worked with the application. For example, a researcher may ask individuals about their own expectations of how sound would work in a specific case, or what type of experience the participant has had with sound applications. In this case, the researcher wants to make sure that the participant is not biased by exposure to the target application. During the interaction with the application, participants may be required to provide specific responses that they might forget by the end of the session. However, most of the time, participants complete surveys after they have completed their interaction with the sound application. In this case, the survey serves the purpose of measuring the overall reactions to the application.

The questions in a survey are dictated by the particular application and concerns of the researchers. Demographic questions concerning age, gender, and other relevant personal characteristics should be selected carefully. For example, a researcher may find that women have a preference for a particular type of sound while men may prefer another. Obviously, this would be good to know and could result in the auditory display offering a variety of sounds in a “sound palette” (see Bonebright & Nees, 2008 for an example) to provide the best possible match of the application with the widest possible user audience. General questions about annoyance and distraction levels, overall satisfaction with the user interface, and whether the participant would use such a product would be particularly pertinent for sound applications. Finally, questions that are specific to the target application should be carefully prepared and selected to make sure that researchers have the information they desire. It should be strongly emphasized that construction of surveys can appear deceptively simple to someone who is uninitiated into this type of research. However, effectively wording questions for surveys takes experience and careful consideration of the target population. In addition to the obvious need of writing the questions clearly, using vocabulary that is familiar to the participants, and keeping questions as short as possible, survey items also need to be constructed to avoid leading or biasing questions. One common pitfall researchers make is to construct a questionnaire that is excessive in length. This should be avoided by carefully choosing items that will provide the necessary information to promote the design of the auditory display.

Responses to surveys can take a number of fixed response format items, such as rating scales, true or false questions, and check boxes for relevant properties, as well as free response options. However, particularly for the purposes of evaluation to guide design or refinement of a product, a good general guideline is to make more use of rating scale questions (e.g., 5, 7, or 9-point scales) rather than yes/no questions. Data from fixed response format items are easier to analyze, but free responses may provide a richer source of data. In many cases, a combination of fixed and open response items may provide the best balance for both the researcher’s purpose and the ability of the participants to respond in a way that reflects their true opinions.

Surveys provide a relatively easy way to determine users’ opinions about auditory displays, but they are not without shortcomings. For example users may react to the perceived demand characteristics of the research context, or they may also respond in ways that they believe are socially desirable. In both cases, the data provided by the participants does not reflect their true opinions or experiences and will lead to erroneous decisions about the effectiveness of the display. (For a good general reference for survey design and construction, see Bradburn,
6.6.3 Verbal Protocols

Verbal protocols require subjects to talk aloud while they work with an application. The participants’ statements can be recorded and/or an experimenter can take notes and cue the participants to elaborate on their comments during the session. The advantages of this type of procedure are that participants do not need to rely on memory in order to report their responses at a later time and that participants can provide spontaneous comments about improvements or problems while they are working with the application. Some researchers have pairs of participants work together since this leads to more information for the researcher while the users explain aspects of the program to one another (Schneiderman, 1998). This approach may in fact lead to a more realistic evaluation of a sound application in two ways. First, when people learn a new application, many times they will have someone help them. Second, it could be especially informative for sound designers to determine whether the sound helps or hinders in what can be a social process.

In spite of the possible advantages of using verbal protocols to evaluate use of auditory displays, there are also potential disadvantages that should be considered before adopting this method. Some of these issues are general problems encountered with use of verbal protocols for evaluation of any product or process, while others are unique to (or perhaps even exacerbated by) situations involving evaluation of auditory displays. One general problem is often encountered when recording sessions by electronic means, such as videotaping or use of digital recording media. Use of passive recording methods can be falsely reassuring since a novice researcher will assume that this means that there is a permanent record of all aspects of the session. Unfortunately, the reality of using recording media is quite different from that expectation. For example, the verbal record can become obscured when the participant doesn’t talk loudly enough, or the camera may be placed in such a way that there are important details that are not captured on the recording. It should also be noted that the recorded media will need to be coded at some point for analysis purposes, and while the researcher can choose to replay a section that was missed, the coding stage will still need to be completed at a later time than if it were done while the participant was interacting with the application. However, if the researcher chooses to have the session recorded by an investigator, it is extremely important to make sure that there is sufficient training so that investigators are consistent across sessions themselves and show a high degree of consistency with any other investigators working on the project. Finally, when examining the effectiveness of an auditory display with the participant talking about the experience, the researcher needs to be aware of any system sounds that might be missed due to the monologue of the participant.

It is important to note that verbal protocols were developed primarily for evaluation of computer software during usability studies (Virzi, 1992). To date, there has been limited use of this technique for auditory displays; therefore, researchers interested in trying this technique should keep in mind that the verbal protocol in addition to listening to an auditory task may have much larger effects on cognitive load and resource allocation than are seen when this technique is used for visual or text based scenarios.
6.6.4 Focus Groups

In focus groups, participants assemble with a discussion leader to provide reactions and opinions about a product (in this case, a sound application) that is being developed. The discussion leader typically has a list of items that he or she wishes to use as beginning discussion points. Such a list is normally generated by the researchers prior to the meeting to illuminate any of the facets they are considering trying or testing. However, it is also important to leave the conversation open so that the participants can bring up issues that are important to them that the researchers may not have anticipated. When working on sound applications, it would be most likely that focus groups would be conducted when the design was for a specific population, such as firefighters or physicians. In these cases, the researcher can gain valuable insight into the needs of the specific group that can then be considered during the subsequent design process. When conversation goes dry, prompts must not be leading so that the conversation is not biased toward a particular topic. Thus it is very important that discussion leaders be carefully trained.

Focus groups tend to consist of five to six participants so that there are not so many people that individuals become bored waiting for their turn nor that there are so few that there aren’t enough voices to keep up the synergy. Individuals chosen for the group should also be carefully selected to make sure all constituents of the target group are involved. Finally, group dynamics must be managed well by the leader in order to get good information from all members.

Analysis of data from focus groups typically involves content analysis of the topics. In most cases, the discussion leader records the major points, as well as the emphasis placed on each, on a checklist type of format that leaves room to specify the topic and provide a rating of significance for the group. Once the data are collected, the content is analyzed for overlapping themes that can help further development of the display. Electronic methods of recording focus groups can also be used to assist with these analyses. However, some of the same caveats presented in the discussion of verbal protocols about the use of recorded media apply here as well. (See Jordan, 1998 or O’Donnell, Scobie, & Baxter, 1991 for further discussion of focus groups.)

6.6.5 Expert Appraisals

Enlisting the help of experts in relevant areas in which designers or researchers are not trained can and should be used when developing auditory displays. For example, a professional who works in sound synthesis will not necessarily have the expertise to make appropriate judgments about the human physical and cognitive limitations that are important to take into account when building a sound application. Or a researcher designing an application for visually impaired people will not necessarily be an expert on the types of needs of this particular population. In such cases an expert appraisal performed by a professional in the appropriate field can be used effectively to avoid pitfalls and streamline the entire process. One way an expert may perform an appraisal is to use a checklist when evaluating the proposed design for an auditory display. An example of this would be to have a perceptual psychologist check on a number of the known perimeters that can affect people’s ability to use a display. It has also been shown in a number of usability studies that multiple experts contributing to the evaluation in their area of expertise can increase the benefits of this
technique (Jeffries, Miller, Wharton, & Uyeda, 1991; Karat, Campbell, & Fiegel, 1992). This approach is particularly useful in the beginning stages of the project even before a prototype is built, but it should be used with other methods of obtaining information from the target population as mentioned previously in this section.

6.7 Conclusion

In concluding this chapter, it may be helpful to summarize some of the most important global “take home” messages:

- Researchers in auditory display need to use appropriate auditory design principles and good research methodology.
- Good design projects will likely use multiple research methods to provide sufficient information to produce a good display.
- The context for the auditory display and the target population must be included in the design of the display from the beginning of the process.
- Researchers should seriously consider using teams that include individuals with complementary training to assure that all aspects of auditory design are addressed.
- Human auditory perceptual abilities must be a central consideration for the development of auditory displays.
- Statistical techniques should be used when appropriate but should not replace real-world testing nor mitigate the practical significance of sound applications.
- Decisions about the appropriate statistics to use must take into account the ultimate goals of the project.

Finally, the authors wish to note that researchers working in the development of auditory displays have made great strides in applying appropriate techniques for evaluating the usefulness of such applications in a variety of contexts. Hopefully this chapter will further facilitate extension of these techniques into the discipline and will act as a catalyst and reference for both experienced and new researchers in this area.

Bibliography


