# Subjective Evaluations of Acoustic Sources

by

Nathan Eyring

Submitted to Brigham Young University in partial fulfillment

of graduation requirements for University Honors

Department of Physics and Astronomy

Brigham Young University

July 2015

Advisor: Timothy Leishman

Honors Representative: Lawrence Rees

Signature: _____ Signature: _____

# ABSTRACT

Subjective Evaluations of Acoustic Sources

Nathan Eyring

Department of Physics and Astronomy

Bachelor of Science

The main purpose of this thesis is to provide an introduction into the methodology of performing subjective analysis of acoustic sources. The author has found that it is not uncommon for those with a technical background to overlook the unique challenges associated with subjective assessments. The thesis describes the ethical procedures of working with human participants and gives aid in creating an IRB proposal. It then provides a brief introduction into how to perform statistical analysis. Finally, it discusses how to develop, evaluate, and implement a subjective evaluation. The appendixes provide examples of the author's work in this area. It includes a listener training program that was developed and implemented at BYU, along with an analysis of the results. It also includes a subjective evaluation performed to evaluate the impact of new directivity data in room acoustic modeling software.

ACKNOWLEDGMENTS

I wish to give a special thanks to my advisors who supported me in a cross-disciplinary project: Dr. Timothy Leishman who appreciated and encouraged research in an area not strictly related to physics and Dr. Bruce Brown who guided me into the world of Psychoacoustics. I also wish to thank my fellow students who helped me complete the work. In particular, I wish to thank Zac Jensen who enabled me to set up and take measurements in the De Jong Concert Hall, Jenny Whiting who developed the model of the hall, Jay Eyring who gave me something to test, and Hillary Jones who was crucial in the administration of the listener training. This thesis was only possible due to the great aid I received from those around me.

# Contents

# Chapter 1

# Introduction

The audio industry has long claimed excellence of many of its products. However, a legitimate question is this: how trustworthy is its opinion? In the 45th anniversary issue of *Stereophile* magazine, its creator is quoted in an interview saying, "As far as the real world is concerned, high-end audio lost its credibility during the 1980s, when it flatly refused to submit to the kind of basic honest controls (double-blind testing, for example) that had legitimized every other serious scientific endeavor since Pascal. [This refusal] is a source of endless derisive amusement among rational people and of perpetual embarrassment for me, because I am associated by so many people with the mess my disciples made of spreading my gospel" [1]. Another expert in the field, Floyd Toole, stated that listening tests done by most reviewers "violate the most basic rule of good practice for eliminating bias" [2]. And this lack of proper testing in the world of electroacoustics has clearly had its consequences. Toole showed that even some high-end loudspeakers, which cost a small fortune, have very low sound quality [2].

If this is the case, it causes one to wonder how one might properly measure the quality of such products. In the world of electroacoustics there are numerous objective and computational methods that attempt to measure quality with varying levels of success. These

measures still need to be corroborated by the opinions of those using the products. What does it matter if a computational model predicts loudspeaker A will sound better than loudspeaker B if anyone who listens claims that loudspeaker B is the superior product? It is therefore crucial to have a basic understanding of how to perform fundamental subjective analysis of these devices.

That is the motivation of this thesis. There are countless books on the topic of scientific subjective research methods, but the vast majority are focused toward students of the social sciences. Thus, these books cover many topics not pertinent to measuring electroacoustic devices. The few books focused on this subjective assessment are also broad. While many of these resources are excellent, they may prove difficult for a simple introduction into the world of subjective analysis.

This thesis is intended to serve as a manual for those with a technical acoustic background and little knowledge of subjective analysis. It will provide enough understanding to enable one in the field to venture into the new realm of subjective acoustic research with the right questions to develop research appropriately. This is not to say that it will be sufficient to provide everything needed to develop any research. However, despite their limited exposure to subjective testing methods, acoustics students of Brigham Young University can develop, implement, and analyze basic psychoacoustic evaluations. This honors thesis provides the fundamental concepts to familiarize them with the tools to do this.

## 1.1 About This Thesis

This manual has three main sections. The first introduces the Institutional Review Board, the organization that oversees ethical considerations when dealing with human participants. It also gives suggestions on what to consider when preparing a proposal to perform research

on human participants. The second section gives a brief introduction into statistical analysis and outlines a few key statistical methods. It gives enough tools to do an initial analysis of data. The third section discusses test design, evaluation, and implementation. It gives enough understanding of test design and implementation that someone with no experience with subjective evaluations can design an appropriate test and implement it in such a way that the data being gathered is valid. Appendix A describes a listener training program developed at BYU and Appendix B describes research looking at the subjective quality of a new style of directivity measurements using the ideas discussed in this thesis.

# Chapter 2

# Institutional Review Board (IRB)

When performing research involving human participants, there are many ethical considerations of which one should be aware. Every research institution is required to ensure all researchers properly implement these considerations. To do this, each institution has something called an Institutional Review Board (IRB). In order to perform research on human participants, the IRB must give its approval. This is done most often through review of an IRB proposal. This section will describe what the IRB will be looking for when approving research and what must be included. It will describe the background which was provided in the Belmont Report [3], and then give specific suggestions regarding the do's and do not's when preparing an IRB proposal.

## 2.1 Belmont Report

To protect human subjects, the National Research Act was passed into law in 1974. This act commissioned an investigation into basic ethical principles that govern biomedical and behavioral research. The Belmont Report summarizes this investigation. In order to perform research with human subjects properly, one must understand the concepts of this report.

The pertinent details will be covered here as a basis. Nonetheless, if the reader has further questions on the ethical basis of biomedical and behavioral research, he or she should consult the report.

The report has three main parts: boundaries between practice and research, basic ethical principles, and applications. A discussion of these with an acoustical perspective is important.

### 2.1.1 Boundaries Between Practice and Research

The National Research Act does not apply to every project involving humans, especially in acoustics. The law distinguishes between practices and research. "The term 'practice' refers to interventions that are designed solely to enhance the well-being of an individual patient or client and that have a reasonable expectation of success"[1]. As an example, performing some type of medial treatment to aid someone's hearing would be a practice. If a project is only practice and does not include research, then the National Research Act does not cover the human participants.

Research "designates an activity designed to test a hypothesis, permit conclusions to be drawn, and thereby to develop or contribute to generalizable knowledge (expressed, for example, in theories, principles, and statements of relationships)." Research on human participants is what the IRB regulates. Some evaluations of loudspeaker quality, according to this definition, are not research. For instance, when a group of professional listeners are hired to perform an evaluation of a loudspeaker, if no research is being done on the listeners IRB approval may not be required. Also, as research pertains to generalizable knowledge, many projects performed by an acoustical engineer for commercial reasons are not research. A company may "research" their own products for their own evaluation, but if

[1] All quotations in this section are from the Belmont Report [3]

the information they gain does not contribute to generalizable knowledge, it does not count as research according to the National Research Act[2].

Understanding what counts as research of human participants, according to the Belmont Report, is crucial. Any project that does not fall under the National Research Act does not require IRB approval. The research is thus saved from countless hours of associated work. As a result, early on in any project, one should determine if the project counts as research of human participants. If there is any doubt as to whether IRB approval is needed, each IRB has an office and administrator that can be contacted to discuss the matter.

### 2.1.2 Basic Ethical Principles

The next section of the Belmont Report covers the "general judgments that serve as a basic justification for the many particular ethical prescriptions and evaluations of human actions." It considers these judgements using these principles: respect of persons, beneficence, and justice.

#### Respect for Persons

Respect for persons has two main points: treat individuals as "autonomous agents" and those with diminished autonomy should have protection. An autonomous person is "an individual capable of deliberation about personal goals and of acting under the direction of such deliberation." This means all participants must be able to choose what they want to have happen to them. In order to choose, participants have to have the opportunity to choose what does happen and they themselves must be capable of making that choice. Someone

[2]As another example, in a teaching situation, a teacher may perform analysis on the results of his or her students's work and grades without needing IRB permission if the teacher does not try to generalize the data and report it to the general public.

with diminished capacity of making a choice in regards to the research has diminished autonomy. It is general policy to keep those with diminished autonomy from participating in research unless necessary, and when necessary those with diminished autonomy should have a representative to help them make that choice.

### Beneficence

In research with human participants, researchers must secure their well-being. A two rule summary is: "(1) do not harm and (2) maximize possible benefits and minimize possible harms." These rules apply to both the individual participants and to general society. This means that when contemplating research that does contain possible risks - physical or emotional - to participants, it is important to consider what benefits the individual or society would gain. If there is a sufficient benefit to the participant or society, the research may be justified. Nonetheless, researchers should always minimize any risk to the participant and should not commence research until they have taken every reasonable measure to protect them.

This frequently occurs in cases where researchers use deception in the research. Deception limits the autonomy of individuals, because they do not fully have the opportunity to choose what would happen to or around them. However, some research absolutely requires deception. If the deception has minimal risks, that research often still is approved.

### Justice

As research with human subjects requires beneficence, justice concerns how this beneficence is distributed. It is not ethical to benefit one part of society at the cost of another. To define justice there have been some basic formulations made: "(1) to each person an equal share, (2) to each person according to individual need, (3) to each person according

to individual effort, (4) to each person according to societal contributions, and (5) to each person according to merit."

When considering each of these formulations, justice applies to a vast range of applications in research, including human participants. From selection of participants and development of the research, through its execution and analysis, to the publishing of the data and distribution of the results, researchers must consider justice. The specifics are covered in the next section.

### 2.1.3 Applications

When seeking approval from the IRB, these concepts form the basis of whether permission will or will not be granted. As such, it is crucial to understand how these principles apply to the research at hand. The Belmont Report uses the examples of informed consent, assessment of risks and benefits, and selection of subjects to illustrate the application of respect for persons, beneficence, and justice respectively. This section discusses these examples to illustrate how to apply the basic ethical principles while creating an IRB proposal.

#### Informed Consent

"Respect for persons requires that subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied." The Belmont report divides these adequate standards into three elements: information, comprehension, and voluntariness.

**Information** Research participants should be informed of what will be expected of them, possible risks/discomforts, possible benefits to them and to society, confidentiality, com-

pensation, and an explanation that their participation is voluntary and that they may withdraw from the research at any point. Most often, researchers gain this through a written consent form that participants in most cases will sign. In cases of very low risk (such as an online survey), many IRBs have resources such as a template available to help provide adequate information to the participant. In cases where deception is used, researchers obviously do not include the deception in the informed consent, but they must explain the deception in a subsequent debriefing and have the IRB's approval.

**Comprehension** If the researcher tells the participant every aspect of the research but the latter does not comprehend what the research requires - either through fault of the researcher explaining or the participant's lack of capability to understand - the participant's autonomy has been compromised. This means that all materials should be very clear and simple. They should be written with rarely more than a middle-school vocabulary, and when there are participants who do not speak English fluently, the informed consent should be in their language. It is the researcher's responsibility to ensure the comprehension of each participant, not the participant's. In cases that the participant by himself or herself may not be capable of understanding, a representative for that participant that comprehends both the research and the participant's needs should help decide if the participant should participate.

**Voluntariness** Participants should agree on their own terms to consent, and should not be pressured into participating. This applies to many areas of an IRB proposal. Compensation should not be so great that participants feel the necessity to do things they would otherwise be unwilling to do because of financial distress. Professors should not pool from their own students as his or her students may feel obligated to perform the research. People should not be manipulated into performing the research. A researcher must be wise when considering

if the research protocol truly allows participants to volunteer themselves to the research at hand.

### 2.1.4 Assessment of Risks and Benefits

When the IRB board reads a proposal, their primary task is to assess if the risks presented to the participant are justified given the potential benefits. When potential participants have informed consent they will weigh the risks and benefits and decide if they will participate. As a result, it is very important to fully understand what risks and benefits are part of any research being performed.

To convince the IRB board to grant permission for research, "benefits and risks must be 'balanced' and shown to be 'in a favorable ratio.'" In most cases, participants will not receive direct benefits from the research performed, which the researcher should state. However, beneficence also considers potential benefits to general society. Quite often, IRB proposals carefully outline potential risks to participants and describe potential benefits to society. Any risk and/or benefit for the participant, participant's family, or society is part of the consideration.

In order to predict risks and benefits, any proposed project must be well considered and defined. This is proper technique in any research field regardless of involving human participants. Prior to the beginning core work, researchers should investigate the plausibility of the anticipated results. Reasonable predictions of how participants will react should also be clear. If there is prior research to draw from, it helps define the risks and benefits.

Finally, the Belmont report gives a list of items that researchers should consider in every project:

> (i) Brutal or inhumane treatment of human subjects is never morally justified. (ii) Risks should be reduced to those necessary to achieve the research

> objective. It should be determined whether it is in fact necessary to use human subjects at all. Risk can perhaps never be entirely eliminated, but it can often be reduced by careful attention to alternative procedures. (iii) When research involves significant risk of serious impairment, review committees should be extraordinarily insistent on the justification of the risk (looking usually to the likelihood of the benefit of the subject - or, in some rare cases, to the manifest voluntariness of the participation). (iv) When vulnerable populations are involved in research, the appropriateness of involving them should itself be demonstrated. A number of variables go into such judgement, including the nature and degree of risk, the condition of the particular population involved, and the nature and level of the anticipated benefits. (v) Relevant risks and benefits must be thoroughly arrayed in documents and procedures used in the informed consent process.

### 2.1.5 Selection of Subjects

Justice largely determines where participants are pooled from and what they are used for. This may appear on both a social and an individual level. On an individual level, if some benefit is offered to one group of participants and is not offered to another group (such as a control), it is unjust that all those performing the same work receive different benefits. For example, if researchers are measuring the aid of vocal training, it is unfair to select one group of participants to receive vocal training and one group to go without, if both are being measured for their vocal health. There is a simple remedy, however. Researchers can offer the control group whatever benefit they might have received following the completion of the research or some other form of equal compensation.

Researchers should also consider the selection of subjects on a social level. For ex-

ample, it would not be just to recruit participants from a specific demographic (be it race, gender, or cultural background) for a project that will aid some other demographic. Justice dictates that the selection should be representative of the demographic the research is attempting to represent. It is also important to consider those in society that are already in a burdened state. Selection of a burdened group for research that places them at the risk may be unjust, depending upon the project.

## 2.2 Further Applications

The Belmont Report does explain some of its important applications and gives appropriate examples. However, there are many individual considerations pertinent to IRB proposals that are not discussed in the report. Accordingly, there is much more needed that should be considered to build an IRB proposal. Most IRBs have an outline form on their website or in their office that should be carefully followed. This section will not outline every aspect of writing an IRB proposal, but many individual details herein are important to remember. The section will provide a checklist of items to consider when preparing a proposal. It is not a comprehensive list, but it serves as a basic guide. The materials provided by each institution's IRB should be the primary resource.

**Assessing Whether an IRB Approval is Necessary**

- Gaining IRB approval requires a lot of time and is not needed in many acoustics applications. If the research fulfils the following list, it may not need IRB approval:
  - There is minimal risk to human participants.
  - The project's purpose does not include researching the human participants (according to the definition of research given in the Belmont Report). If loud-

speaker quality is the only consideration, IRB approval may not be necessary.

  - Participants themselves would feel as if they are not the focus of the study.

- If there is any question as to whether IRB approval is needed, most IRBs have an administrator or secretary that can be contacted for clarification.

**Researchers**

- The principal investigator listed on the IRB proposal does not necessarily need to be the principal investigator of all research done. He or she is simply the person responsible to cover the IRB protocol for the project.

- Most IRB boards require all research personnel to have done some sort of training on IRB protocol. One of the most common is called CITI training (easily found on the websites of IRBs that require them or by using an online search engine).

- When describing qualifications of research personnel, only enough information to ensure they are capable of performing the research without risk to the participants is needed. For minimal-risk projects, this may be one or two paragraphs (assuming they have done the required IRB protocol training).

**Participants**

- Use vulnerable populations only if needed. These include but are not limited to: children, pregnant women, prisoners, economically disadvantaged persons, institutionalized, students enrolled in the classes of the researchers, mentally disabled persons (an individual who is unable to provide informed consent for himself or herself), and/or educationally disadvantaged persons.

- Without knowing it, it is often possible to recruit someone who is part of a vulnerable population. If it is reasonable that the researcher does not know the participant is part of a vulnerable population and the research poses no known added risk to them, research protocol does not have to adapt for this. (For example, research protocol does not have to include investigation to know if someone is economically disadvantaged if economic standing has little impact on the research).

**Confidentiality**

- Any data that contains any personal information about participants (including information that would indicate they participated in the research) must be secured by lock and key for paper sources and password protection for electronic sources.
- If not needed, do not gather identification information from participants. If it is necessary, describe the details of how this information will be secured.
- If participants are identifiable in any publication, permission (usually written) is needed from them (this can be gained in the consent document).
- Any personal information obtained about participants must be justified.

**General Tone**

- The IRB proposal should lay out possible benefits and risks, and be able to convince the IRB that the research is necessary.
- When describing the planned research to be done, it is often wise to overstate how much will be done if overstating does not add more risk to the participants. The IRB board will not punish a group for not performing every research protocol described, but they will stop research that is not in the proposal.

- The IRB proposal may ask for seemingly specific details such as the number and ages of participants, analysis strategy, potential risks and benefits, etc. This is to ensure that there is proper preparation, but is not restrictive. A different number of participants and/or analysis techniques may be used than what was stated. All details on treatment of participants must be included and is resctrictive.

- Compensation, when given, should be merely sufficient to compensate for the participant's time and effort. It should not be so much as to compel someone to do the research. If different participants receive different amounts of compensation, the researcher should explain the reasoning and qualifications for differing compensation levels.

- Compensation is not a benefit and should not be called that in the IRB proposal.

**Informed Consent Document**

- The language used should be simple. Every participant must be able to understand it.

- The document should cover the basics of the research procedures, risks, benefits, confidentiality procedures, and compensation as described in the IRB proposal.

- Give an estimate of the time commitment.

- Indicate in some way that participants may choose to end their participation in the research without any risk or consequence to them.

- Give the contact information of someone who can provide information should the participant have any questions.

**Background Research**

- Cite any test, survey, or any other portion of a project based on old research. An already published process often helps gain IRB approval.
- The background section's purpose is more or less to see if the researchers have done sufficient background research to understand potential risks and benefits. As such, it should include multiple applicable references and a sound discussion of how they relate to the research project at hand.

**Deception**

- Do not use deception unless necessary.
- The IRB proposal should be clear as to why any deception is necessary.
- Describe the debriefing and the explanation of the deception. If the debriefing does not inform participants about the deception, give a justification.

## 2.3 Conclusion

When working with human participants, the researcher must take extra precautions. Proper research technique is not just a good idea, but also a necessity to ensure the safety of those participants. This might be cumbersome when something like an IRB proposal must be written, but knowing the rights of each participant helps to keep the research effective.

# Chapter 3

# Statistical Analysis

Once data has been gathered, it is crucial to understand how to properly interpret it. Even if every aspect of the test design is correct, conclusions that do not include a proper analysis may be incorrect. As stated by Bech:

> The importance of detailed statistical planning of the experiment and subsequent analysis of the results is often not acknowledged by audio engineers (nor practitioners from other disciplines as well). The physical variables (the test set-up, the stimuli, etc.) are usually controlled to an acceptable degree; however, the results are often only reported by simple mean values without confidence intervals, for example. Such a lack of statistical information may lead to incorrect scientific or application decisions.

Chapter 4 describes proper test design, evaluation, and implementation and covers correct planning. This chapter will cover proper statistical analysis of collected data. This will include a brief description of the theory behind the statistics and a description of a handful of useful statistical methods. While there are many other statistical methods, the following will give enough detail for basic analyses.

## 3.1 Theory of Statistical Analysis

In scientific testing there is a basic structure: in a given system some variable will be controlled by the researcher (the independent variable) and as the researcher changes the independent variable he or she will measure how another variable changes (the dependent variable). The goal is then to predict how the dependent variable changes with respect to the independent variable. In subjective research, this may relate to a researcher comparing different loudspeakers (the independent variable) by playing each of them to see how their subjective quality is rated by listeners (the dependent variable). Generally, the researcher's goal is that the average response of the group or sample predicts the average response of all those interested (the population) in using that loudspeaker.

However, there may not be any connection between the independent variable and the dependent variable. It is possible that two loudspeakers sound nearly identical and so changing the loudspeakers results in no difference in the perceived quality by the participants. In statistics, the assumption is that there is no connection between the independent variable and the dependent variable. This is the null hypothesis.[1] As the point of the research is to find the connection between the independent and dependent variables, statistical analysis is built to show when it is reasonable to reject the null hypothesis and apply the differences found by the measured group to the general population.

Statistical analysis has a chance of giving erroneous results. It might reject the null hypothesis, even when the null hypothesis is true and there really is no connection between the independent and dependent variables. This is called a type I error. The inverse situation may also occur. That is to say the statistical analysis may fail to reject the null hypothesis,

[1] Social scientists will assume the null hypothesis when performing statistical measures in part to limit their bias in the results. The most important aspect of the null hypothesis for one with a technical background is that it is a very common part of the vocabulary in statistical analysis.

even when the null hypothesis is incorrect. This is a type II error. These errors may happen for multiple reasons. Common culprits are insufficient numbers of participants, measuring a sample that is not representative of the population, or an incorrect experimental set-up that introduces unknown variables.

Many statistical analyses report a $p$-value. This is simply the chance on a scale of 0 to 1 that the null hypothesis is true. A $p$-value of 1 indicates that there is a 100% chance that the null hypothesis is true and there is no connection between the independent and dependent variable. A $p$-value of 0.05 signifies that there is a 5% chance there is no connection (or a 95% chance there is a connection). This connection is whatever the statistical method is trying to measure. In subjective research, the commonly accepted maximum $p$-value is 0.05. Anything below this is statistically significant. The lower the value the better.

## 3.2 Statistical Methods

While the theory behind statistics is important to understand, it is necessary in practice to perform various calculations in order to understand varying results. This section will cover a handful of useful statistical methods that will be sufficient to give good insight into the meaning of results. Its purpose is to help build an understanding of the methods and how to use them. Equations will be given to build intuition of what each method is truly measuring. Many software packages can compute these and other statistical methods, including MATLAB and Mathematica.

### 3.2.1 Comparison of Means

The most common goal in research is to see how two groups differ from each other. For example, when comparing the quality of two loudspeakers, the researcher will want to

see how participants viewed each speaker differently from the other. This is done most often by comparing the means of the responses given by participants. For example, if the survey asks how clear each loudspeaker is on a scale of 1 to 10, a researcher can compute and compare the mean value of each loudspeaker's clarity. Using their mean values is a measure of central tendency. As discussed in the theory section, these differences may be due to chance. An inspection of the measure of spread, or a measure of how spread out the values of a set are is also important. As such there are many statistical methods that further illuminate and define measures of central tendency and spread.

**Standard Deviation** One of the most common statistical metrics is the standard deviation. The standard deviation is essentially a measure of how much the values used to compute the mean vary from that mean, so it is a very basic measure of spread. The equation to compute the unbiased sample standard deviation is:

$$\sigma = \sqrt{\frac{\Sigma_{n=1}^{N}(x_n - \bar{x})^2}{N-1}} \tag{3.1}$$

where $x$ is each individual value, $\bar{x}$ is this mean value, and $N$ is the number of values being averaged. In the rare case that the entire population is measured, this equation is divided by $N$ instead of $N-1$. As can be seen from this equation, the standard deviation is essentially the root mean square of the deviation of values from the mean. When a mean is reported, its standard deviation should always be included. The standard deviation gives the most basic reference point to how exclusive the mean is; any value within a standard deviation is somewhat similar to that mean.

***Z* Scores (Standard Scores)** Very closely related to the standard deviation are $Z$ scores, which essentially measure how much a given score deviates from the mean relative to a standard deviation. For example, IQ tests have a mean of 100 and a standard deviation of

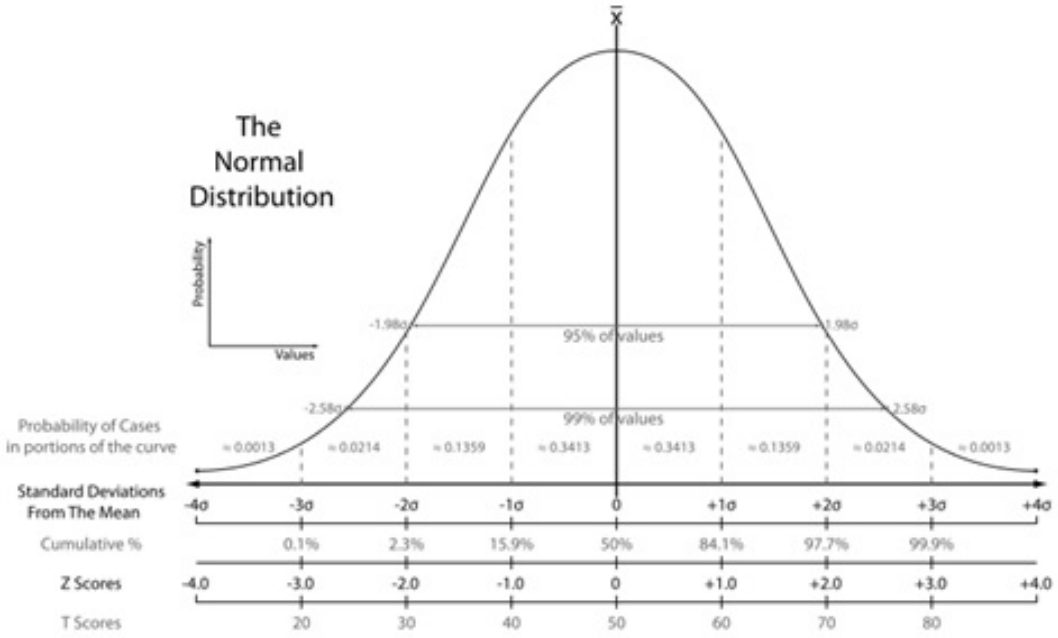


**Figure 3.1** A normal distribution that also shows relations of standard deviation, cumulative percent, Z scores, and t scores. [4]

15, so a value of 115 has a $Z$ score of 1, 130 has a $Z$ score of 2, 85 has a $Z$ score of -1, etc. It is calculated for a given value of $x$ as:

$$z = \frac{x - \bar{x}}{\sigma} \tag{3.2}$$

These scores can be very useful when looking at how different certain values are. When looking at a normal distribution it is quickly and easily possible to predict the likelihood of how similar one value is to the sample using the Z score. Using Figure 3.1, one can see that a $Z$ score of 1 is larger than 84.1% of other values that made up the distribution of scores, and a Z score of 2 is greater than 97.7% of other potential values. While $Z$ scores are not the most common tool used, they are clearly powerful.

***t* Test** The methods derived thus far only compare one value to the mean of some group. It is often very useful to compare two means to each other. In some cases, it is useful

to compare the mean of the control group (with no change to the independent variable) and the group with a change to the independent variable. While their standard deviations should be calculated to see how wide their value distribution is, this alone does not ensure the two values are significantly different and that the null hypothesis should be rejected. To compare two means, a t test is used. The formula for the *t*-value is:

$$t = \frac{\bar{A_1} - \bar{B_2}}{\sqrt{\sigma_A^2 + \sigma_B^2}\sqrt{\frac{1}{n}}} \tag{3.3}$$

where $A$ and $B$ represent each mean being compared, $\sigma_A^2$ and $\sigma_B^2$ represent the squared standard deviations of the two means (or their variances), and $n$ is the number of items in each mean.

To use the *t*-value in a meaningful way, two more bits of information are needed: (1) degrees of freedom (df) and (2) knowledge of whether it is a one-tailed or two-tailed test. Degrees of freedom are calculated by the number of values included in each of the two groups and is $N_1 + N_2 - 1$. If the researcher only cares if one mean is higher than the other mean or only lower than the other mean, than this is a one-tailed test. For instance, if a new loudspeaker design that is being tested with the goal that the sound quality rating is simply no lower than another loudspeaker, a one-tailed test is appropriate. If, however, the researcher only wishes to test whether the two differ regardless of which is larger, then it is a two-tailed test. This would be the case if the researcher is investigating whether a loudspeaker is either better or worse than another loudspeaker

Once all these different parts (*t*-value, degrees of freedom, and if it is a one-tailed or two-tailed test) are known a *t* table can be used to obtain the corresponding *p*-value. A *t* table can be found easily by searching online or by looking in any of a number of statistics text books. A *t* table's rows are labeled by degrees of freedom and its columns are labeled by *p*-values for a one-tailed or two-tailed test. To use a *t* table one finds the row which

most closely corresponds to the degrees of freedom in the research, and then at the head of the column finds the value that most closely corresponds to the $t$-value obtained. In other words, the column that $t$-value is in is labeled by its corresponding $p$-value. When using a computer package, it will calculate the degrees of freedom and $t$-value, but the researcher needs to know if the test is one-tailed or two-tailed.

A significant $p$-value here signifies that the two means are sufficiently different from one another that the null-hypothesis can be rejected. This can be more simply stated by saying the two means are statistically different. It signifies that the two means may be used to compare the two cases in the research. For example, consider two means of data measured on a 1 to 100 scale that are statistically different with mean A being 54 and the mean B being 56. Since they are statistically different, although the two mean values are very close, the relationship of mean B being greater than mean A can be trusted. However, if they are not statistically different (the $p$-value is greater than 0.05), even if mean A is 20 and mean B is 60, the relationship that B is greater than A cannot be trusted.

One other important consideration is that the $p$-values that correspond to $t$-values are based on the assumption that the measured data follows a normal distribution. If the data are plotted in a histogram and do not look like a normal distribution, a type II error is possible. As an example, consider Figure 3.2. One of the means being compared has a normal distribution, but the other is very largely skewed. The means may be very similar for the two graphs, and might have a small $t$-value and corresponding $p$-value. This would fail to reject the null hypothesis, even though there seems to be a very large difference between the two groups. Further analysis should be done to investigate the skewed curve and what creates the pattern. There seems to be something interesting in that data, and simply because the $p$-value is high does not always necessitate that there is nothing statistically different between the two means. Data should always be looked at to see the interactions and potential pitfalls

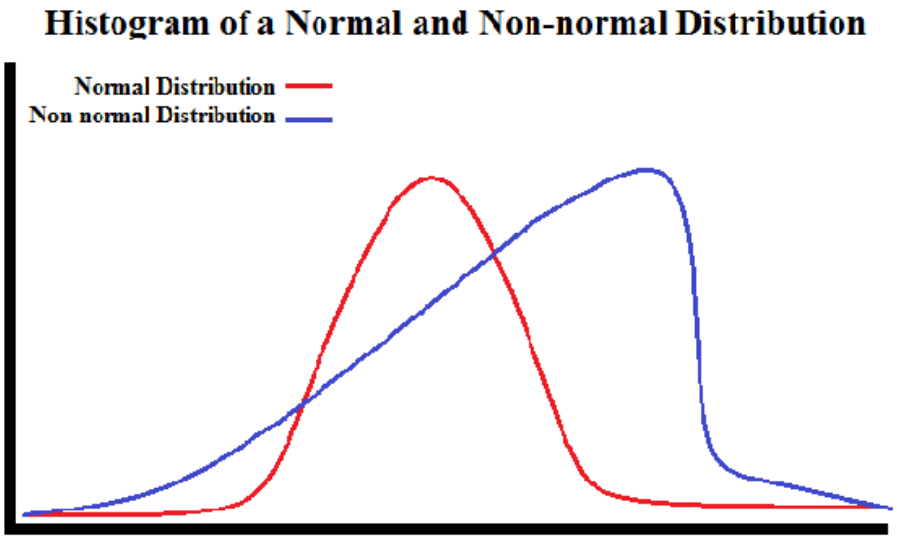


**Figure 3.2** Two sets of values are compared, one with a normal distribution, one with a non-normal distibution. These two sets may give incorrect values when using a *t*-test.

of the mathematical analysis.

**Analysis of Variance (ANOVA)** Often in subjective evaluations, more than one independent variable is used and it is a point of interest to see how many other variables depend on them. For example, when comparing loudspeakers, it is quite common to have as independent variables the multiple loudspeakers being compared, demographic information, and other factors as independent variables. Then the quality, clarity, envelopment, and other attributes may all be dependent variables based on those independent variables. This would result in numerous means being compared. If a *t* test could be used to compare every condition it could become quite cumbersome and eventually lead to false conclusions. Fortunately, there is a tool that compares the means of multiple dependent variables to one independent variable. This is called an analysis of variance (ANOVA). The mathematics to calculate this test will not be described because generally an ANOVA is performed using

a computer software package.[2] A summary of the procedure used to calculate an ANOVA will be given here.

The ANOVA compares the variance *within* each of the individual groups to the variance *between* the means of those several groups. If three loudspeakers are compared, it computes how much the values corresponding to just the first loudspeaker vary, just the second loudspeaker, and just the third loudspeaker. It also calculates how much variance there is between the means of loudspeaker 1, loudspeaker 2, and loudspeaker 3. It then divides the variance between the means of the groups over the variance within the three groups, which gives an $F$-value. Similar to the $t$ test, the $F$-value can be compared to an already existing table to see its corresponding $p$-value. However, the results will most often be found using a computer software package and a $p$-value will be given.

This test is very commonly used in subjective evaluations of acoustic sources. This is because it readily demonstrates whether there are significant differences between acoustic conditions. For example, if multiple loudspeakers are being compared, one ANOVA can see if the type of loudspeaker impacts any dependent variable such as quality, ASW, or clarity.

**Post-hoc Analysis** The simultaneous advantage and disadvantage of the ANOVA is that it considers all changes in the independent variable at the same time. This is good because it is able to measure for statistical levels of variance in means between all the data very quickly. However, it is incomplete because two means may be almost identical but the other means looked at in the ANOVA are different enough to have a statistical difference.

[2] In MATLAB, the command to perform an ANOVA is anova1(y), where each column in y contains the reported values for each dependent variable.

It is inappropiate to perform multiple $t$ tests,[3] so another process is needed. This is done through post-hoc analysis. Post-hoc analysis will look at the general variance between each individual condition and each other condition and often report for each condition its mean, standard deviation, and $p$-value corresponding to the difference in its mean and every other condition's mean.

It is important to note that there are many types of post-hoc analyses. Each has its purpose and is worth studying. One useful method, the Tukey method, assumes each condition is observed independently, that the variation in each condition is nearly equal, and that each comparison is evaluating quantized values. Each individual comparison is very similar to a $t$ test, but corrects for the probability of making a type I error.

This is a very useful test. Most statistical packages can simply be told to calculate it once the data sets are entered.[4] When comparing multiple conditions at the same time, it is often one of the most useful tools in the researcher's tool box.

### 3.2.2 Measures Beyond the Mean

**Pearson's *r*** One other consideration is how much two variables correlate to one another. For example, one can measure how much the clarity of a loudspeaker correlates to the level of its bass content. Using this example, one would determine a series of clarity values that corresponded to a series of specific bass content values (e.g., when the bass content is 10 dB the clarity is 0.1, when the bass content is 20 dB the clarity is 0.15, etc.). Then these

[3] Multiple $t$ tests lead to multiple problems such as alpha inflation, non-independence of comparissons, and other issues. However, these issues are well beyond the scope of this thesis. It simply should be understood this approach is inappropiate.

[4] The command in MATLAB for this test is multcompare(stats), where stats are gained from performing an ANOVA in MATLAB such as [p, table, stats] = anova1(y).

two series would be compared using an equation that returns something called Pearson's $r$:

$$r = \frac{\Sigma_{n=1}^{N} XY - \frac{(\Sigma_{n=1}^{N} X)(\Sigma_{n=1}^{N} Y)}{N}}{\sqrt{(\Sigma_{n=1}^{N} X^2 - \frac{(\Sigma_{n=1}^{N} X)^2}{N})(\Sigma_{n=1}^{N} Y^2 - \frac{(\Sigma_{n=1}^{N} Y)^2}{N})}} \tag{3.4}$$

where $X$ corresponds to one variable (the clarity), $Y$ corresponds to the other variables (the bass content), and $N$ corresponds to the number of values being summed over. Pearson's $r$ says how strongly these two values are correlated on a scale of -1 to 1. A positive value indicates there is a positive correlation (or that as $X$ increases, so does $Y$) and a negative value corresponds to a negative correlation (or that as $X$ increases, $Y$ decreases).[5] It is important to understand that this measures a linear correlation related to a polynomial of degree one between the two sets.

From Pearson's $r$ another important bit of information can be gathered, the proportion of variance in one variable accounted for by the other variable. Proportion of variance accounted for is just $r^2$. It is referred to as the coefficient of determination. The coefficient of determination is a measure of how much the variability of one series can be accounted for by the variability in the other series. So in the example of measuring the correlation between clarity and bass content, if $r = 0.5, r^2 = 0.25$, and 25% of the variability of the clarity can be attributed to the variability of the bass content.

These correlation values are not necessarily statistically significant. So a corresponding $p$-value must be found for the $r$-value as well. This is done quite often through tables (which like the others may be found online). This gives the minimum $r$-value to obtain a certain $p$-value corresponding to a certain level of degrees of freedom (calculated by n-2). It may be a one-tailed or two-tailed test. For example, if there are 20 points of comparison there are 18 degrees of freedom, and if it is a two-tailed test the table indicates that there

[5] The command in MATLAB for this test is corr(X,Y,'name',value), where the name specifies the type of correlation. Pearsons's $r$ is the default correlation method.

needs to be a correlation of at least 0.444 in order for it to be statistically significant at a $p$ = 0.05 level.

This test is very useful because it can show interesting results in the gathered data even when the means by themselves are not statistically significant. It can also help build intuition between how subjective measurements correspond to objective measurements. For example, subjective clarity can be measured and compared to the values of an objective clarity measure such as C50 altered. When the correlation is found between the two, it will be easy to see if the objective C50 measurement corresponds to what people subjectively consider clarity.

## 3.3 Conclusion

When looking at results from subjective research, proper statistical analysis is crucial. If not used, any conclusion based on the data may be invalid and cannot be trusted. The point of this chapter was to give a brief introduction to statistical analysis and give a handful of tools for acoustical researchers to use. To develop further knoweldge of statistical analysis the reader is referred any of a very large number of introductory statistics texts. While the treatment here is not comprehensive, it gives the novice the chance to begin to understand ways in which the gathered data can be approached.

# Chapter 4

# Test Design, Evaluation, and Implementation

## 4.1 Constructing a Subjective Test

Before beginning to create a subjective test, it is very important to understand what the goal of the research is. The initial question generally is very vague and consequently would be difficult to turn into a valid test. For example, consider if the inquiry is how good a loudspeaker sounds. If someone were to ask what "good" means it presently would be difficult to answer. This is because "good" is only an abstract idea, or a hypothetical construct.

In order to perform a measurement based on a hypothetical construct, it is important to give it a precise definition. Quality may mean different things for different audiences, and so it must be operationally clear what quality means when testing for the quality of a loudspeaker. For example, a possible definition of a high-quality loudspeaker is a loudspeaker with wide apparent source width (ASW), neutral color in the sound, and high clarity. This is

the operational definition of the hypothetical construct - or the given definition for quality. This original hypothetical construct is now more specific; however, other potentially ambiguous concepts define the loudspeaker. To solve this problem, the operational definition must be defined further in such a way that they can have non-abstract answers.

How to further define the operational definition will be demonstrated using the three aspects mentioned above (ASW, color of the sound, and clarity). They will be used as specific examples of how questions may be created in order to test a loudspeaker.

### 4.1.1 Simple Question Development: ASW

A simple example relates to ASW. If someone were to listen to a live performance by an orchestra and were to close their eyes, their ear-brain systems would perceive the sound coming from the full width of the stage. The ASW characterizes how wide the sound source appears to be whether the sound source is visible or not. A set of high-quality loudspeakers should cause a participant to hear a wide sound stage in front of them (given a properly mixed and mastered recording), so by asking participants to judge the width of the perceived sound source helps provide data on a quality that is testable.

### 4.1.2 A More Complicated Example: Color

Some of the other concepts in the operational definition may be more difficult to ask the participant. For instance, the idea of a loudspeaker giving a neutral color to the sound is not as simple to define as the ASW. A participant may be confused as to what a neutral colored sound is as opposed to a noticeably colored sound. As the idea of a loudspeaker with neutral color may still be an abstract idea, it is important to design specific questions in order to obtain usable answers to the questions.

To further illustrate the complexity of how a loudspeaker might be colored, a useful

comparison is imagining eating spicy food and someone asks how spicy it is. The response could simply be "very spicy," but someone accustomed to spicy foods might give a more complete answer. The spicy food's hotness may linger longer so that it leaves the mouth burning after eating it. It may have a spicy flavor but not be quite as hot as other spicy foods. It might be Mexican spicy or Indian spicy. All of these many different aspects make up what spicy might mean to an individual. The ability to define spicy in a more complex way is similar to how the coloring of a sound can be defined using many different aspects.

Accordingly, in order to test how a loudspeaker is colored perceptually, one must ask what makes up that color of a sound. When looked at through objective measures, color tends to relate to the spectral content of the sound. Considering color as the preferred spectral content heard by a listener helps lead to proper questions to ask. In the audio world, a sound is often altered in three main frequency ranges: highs, mids, and lows. Most mixing consoles will have three knobs for each channel dedicated to altering these specific frequency bands in order to color the sound to the sound engineer's taste. As this is a common practice in the audio world, these three ranges are also helpful when creating questions to assess the color of a loudspeaker.

The problem is how to word a question to find a listener's preference for spectral ranges. To evaluate the high-frequency content, for example, a researcher could ask a listener either how "tinny" the sound is or simply how much high-frequency content there is. Some listeners will understand what tinny means, some will understand what high-frequency content means, and some will understand neither. In order to decide how to word the question assessing the high-frequency content, it is therefore crucial to consider the audience. If the participants assessing the loudspeakers are a group of sound engineers, it may be appropriate to ask them to assess the highs of the sound, whereas if a group of physicists is asked, it may be more appropriate to ask how balanced the high-frequency auditory content (5 kHz

- 20 kHz) seems to be. For a group of audiophiles, use of the word tinny might be the best option to assess the color of the sound due to high-frequency content. The audience will dictate how the question is worded.

Once a wording for a question is decided upon (for the high-frequency content example, let us use the question: "How tinny is this sound?") it is important to clearly define the key terms to the participant. Even if every listener is an audiophile, each audiophile has a different background. These different backgrounds mean that each listener may give an erroneous answer to a question if it is not clearly explained and defined. Thus, somewhere in the test administration the subject should be given one definition of tinny: "Tinny means having a displeasingly thin, metallic sound," or "Tinny relates to the high-frequency content and means a displeasingly thin sound lacking in resonance." Both of these definitions are true, but might elicit slightly different responses, which means that it is requisite to clearly give the definition to the listener.

Another important aspect of the question is the method used for the measurement. If the point of testing a loudspeaker's quality is to sell it to consumers, a free-response answer might be useful. Glowing remarks of the color of a loudspeaker would possibly be better for business than any numbers. Nevertheless, free-response questions make it difficult to gain numerical data to analyze, which limits the ability of researchers to find aspects to improve or better understand.

In order to create some type of numerical data, it is necessary to use a scale. One of the most common scales for this purpose is the Likert scale. This is often used in opinion surveys that ask participants if they "strongly agree, mildly agree, mildly disagree, or strongly disagree." This is easily converted to something like measuring high-frequency content. If the question is "How much high-frequency content is there?", the response could be given on an adapted Likert scale as shown in Fig. 4.1.

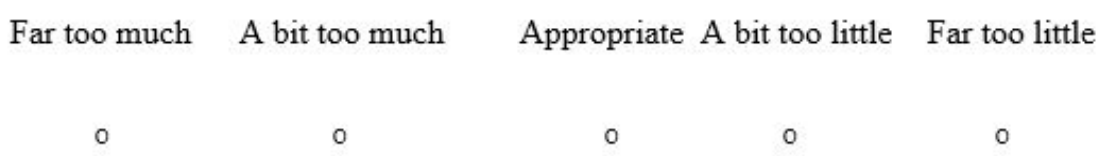


**Figure 4.1** An example of a Likert scale that can be used to measure high-frequency content.

When using a Likert scale, there are a few crucial considerations to keep in mind. First, are there an even or odd number of options? If there are an even number of options, there is no neutral or middle option. This may be helpful if the goal is to force listeners to have either a positive or negative preference. To have a neutral option, an odd number of options is required. Another important consideration is how each option is worded and the relative strengths of their wordings. In order to perform some higher-level statistics it is helpful if the difference between each selection is approximately equal (i.e., the difference between strongly agree to agree and agree to mildly disagree should be equal). One can think of this as asking preference on a -2 to +2 scale. If the wording makes the difference between consecutive options different, the analysis of the results will be faulty.

Another very useful scale is a simple 0-10 scale. For the question "How tinny is this sound?", it is likely the most appropriate scale to use. Tinny is defined as displeasingly thin, and so very few if any will want a tinny sound. This means that a Likert scale would not make sense. In this case, zero has meaning. Zero is the lack of an unpleasant quality. The ten value on the other end may be defined in a definite way to be the maximum value one would possibly give. To increase how definite a ten on the tinny scale is, it is helpful to give a sound that registers as a ten on the scale. Following this example, 0-10 scales are very useful when attempting to measure and compare various aspects of the color of a sound or its quality.

For questions being administered through a computer, sliders are an excellent option

to implement numeric scales. They help represent to the participant a continuous scale as opposed to finite bins. When sliders are used, it is also common to use 0-100 scales in addition to 0-10 scales.

Once there is a question for the high-frequency content, similar questions can be made to look at the mids and lows of the sound. The combination of all three of these questions (high, mid, and low) can give a full answer as to how neutral the color of the loudspeaker is.

### 4.1.3 A Mixed Example: Clarity

Clarity is a potential example of how multiple techniques may be used to look at one aspect. One definition of clarity is "the ease to understand the signal." In this regard, it is possible to make a very simple test similar to what was done with ASW. For example, one could play multiple speech recordings with the speech being more or less clear through a loudspeaker. The number of times the listener is able to correctly repeat what was being said can be defined the clarity score of the loudspeaker. This can be repeated for multiple loudspeakers, and each loudspeaker's clarity scrore can be compared[1].

Clarity can also be defined in a broader sense, similar to how color was defined. Another definition of clarity might be how clear the signal appears to be, how easy it is to understand, and the lack of extra noise in the signal. These three things might all have different answers but make up the perception of clarity. A set of questions and scales can be set up to measure all of these characteristics. The answers to these question can be subsequently put together

[1]If this procedure was used, the researcher must be aware the participant would likely become better at overcoming a loudspeaker's lack in clarity and be able to better repeat the loudspeaker. This means the researcher would need to either have participants measure only one loudspeaker, or have multiple tracks compared by each listener on each loudspeaker. If the latter method is used, proper randomization is needed (see randomization section [Chapter 4.3.1]).

in order to create a full measure of clarity.

### 4.1.4 Putting it All Together

Now there are a series of questions to evaluate different aspects of the loudspeaker. Earlier, a loudspeaker with high quality was defined as one with a wide ASW, neutral color in the sound, and high clarity. Combining all questions should thus enable a researcher to measure and compare loudspeakers' quality. (As an example of a test created using these types of methods, see Appendix B.) This is a relatively simple test. It might not fit the definition of a high-quality loudspeaker for some companies. Further, much of subjective acoustic testing will measure things other than loudspeaker quality. However, this test serves as an example of how to put a test together, and so this can be used as a basic template to create a test for whatever goal a researcher or company may have.

## 4.2 Evaluating the Test

Once a researcher creates a test, it is important for him or her to ask if the test is really any good. When evaluating its quality, there are two major considerations: the reliability and validity. Reliability is how consistent the results are or would be if the test is retaken. If a loudspeaker has high quality today, it should have high quality a week from today and the test should show that. If scores vary a lot and lack consistency, the test does not have reliability. However, reliability does not guarantee that the test is measuring what it was designed to measure. For example, if the test used a poor music selection, it may consistently report one loudspeaker to be better than another, but the loudspeaker the test says is worse may actually be the better loudspeaker. The test is therefore reliable, but not valid (or accurate).

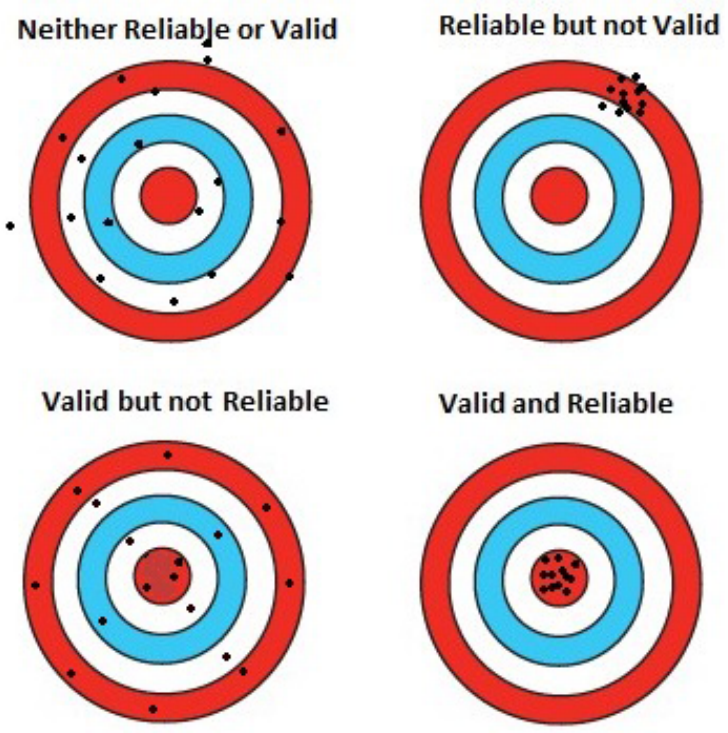


**Figure 4.2** An illustration that shows the difference between validity and reliability. Reliability is analogous to precision, and valdity is analogous to accuracy. Reliability is similar to each throw hitting the same spot, and validity is similar to the average location of every dart thrown being the bullseye.

Validity in a test considers whether the test accurately measures what it reports to be measuring or not. Therefore, if the goal is to measure and compare a loudspeaker's quality, a test has validity if it can accurately and consistently rate loudspeakers. This implies that a valid test should also be reliable. A common comparison involves throwing darts at a target as shown in Fig. 4.2. If the darts hit all over the target, the aim is not precise or accurate. If the darts all hit near the same spot but not the bullseye, the aim is precise but not accurate. If all of the darts hit near the bullseye the aim is both precise and accurate. When measuring the reliability and validity of a test, the reliability is a test's precision and the validity is a test's accuracy.

### 4.2.1 Reliability

To measure the reliability of a test, a common measure is Cronbach's alpha. It measures the internal consistency of each item or question in a test, or the expected correlation between multiple items. It is defined as

$$\alpha = \frac{K}{K-1}(1 - \frac{\Sigma_{i=1}^{K} - 1\sigma_{Y_i}^2}{\sigma_X^2}), \tag{4.1}$$

where $K$ is the number of items included, $\sigma_x^2$ is the standard deviation squared (or the variance) of the total test scores, and $\sigma_{y_i}^2$ is the variance of each item being considered. The alpha value will be between 0 and 1. When inspecting the reliability of a test, this measure indicates how much each item or question in a test is measuring the same thing. For example, perceived spaciousness and reverberation are correlated according to most models, so if an item that asked participants to rate perceived spaciousness and an item that asked participants to rate perceived reverberation are compared, they should have a high alpha value. Generally, alpha values begin to be acceptable at about 0.6, and are excellent around 0.8. A test with an alpha score in that range can be considered reliable. A score of 1 would indicate that every question is measuring the same thing with no variance between the items. As a score of 1 indicates no variance, this is often not desirable because different questions should relate to each other, but provide some unique information that causes them to have a degree of unique variance.

One major concern in subjective testing of acoustic sources, though, is that it may take many factors in order to explain the quality of a sound source. Futher, many of the factors that describe quality may not relate well to each other. For example, the color of a sound is determined mostly by frequency content while ASW is determined by ratios between reflections in the reverberation. So it is possible that color and ASW have very little correlation and would have a low alpha score if compared. In order to check the

reliability of a test considering its multiple factors, another technique must be used as well. This technique is referred to as factor analysis.

Factor analysis processes the combined data of multiple test items to find a set of eigenvalues,[2] which when combined, represent the total variance in the test. In factor analysis, each eigenvalue is considered to be a factor corresponding to the test's variance. From the eigenvectors this analysis shows how much each test correlates to each factor.

To clarify this, a demonstration of how this is done follows from test described in Appendix B. The software Statistical Package for Social Scientists (SPSS) was used which returned the information shown in Fig. 4.3. Each component is a returned eigenvalue from the factor analysis. The magnitude of each eigenvalue was also put into a plot showing their relative heights, commonly referred to as a scree plot (Fig. 4.4). This shows that after a few factors are listed, the slope between eigenvalues begins to be small. To use the fewest number of factors to represent the data set, a cutoff eigenvalue of 1 is generally used. This means that factors with an eigenvalue of more than 1 are used. However, the cutoff may also be found from the scree plot by considering when the slope begins to drastically reduce. Only the eigenvalues above it are used. Another goal is that at least 60% of the variance is explained by the chosen components.

It is then important to understand how each test item relates to each component. In the test explained in Appendix B, there were six test items considered: quality, relative quality between sources (called quality place), ASW, perceived reverberation, how real each file sounded, and how well two sound sources mimicked a live recording made in a concert hall. Each of these items are given an $r$-value to show how well each test item correlates to each component. This is shown in Fig. 4.5. High magnitudes signify that the test item is

[2] A simplified demonstration of the complete mathematics by which one calculates the eigenvalues and eigenvectors of a correlation matrix can be found on pages 149-163 of *Multivariate Analysis for the Biobehavioral and Social Sciences* [5].

Total Variance Explained

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2.437 | 40.609 | 40.609 | 2.437 | 40.609 | 40.609 |
| 2 | 1.240 | 20.670 | 61.279 | 1.240 | 20.670 | 61.279 |
| 3 | .851 | 14.188 | 75.467 | | | |
| 4 | .740 | 12.333 | 87.800 | | | |
| 5 | .451 | 7.519 | 95.319 | | | |
| 6 | .281 | 4.681 | 100.000 | | | |

Extraction Method: Principal Component Analysis.

**Figure 4.3** The values generated by SPSS when a factor analysis was run on the test items described in Appendix B.

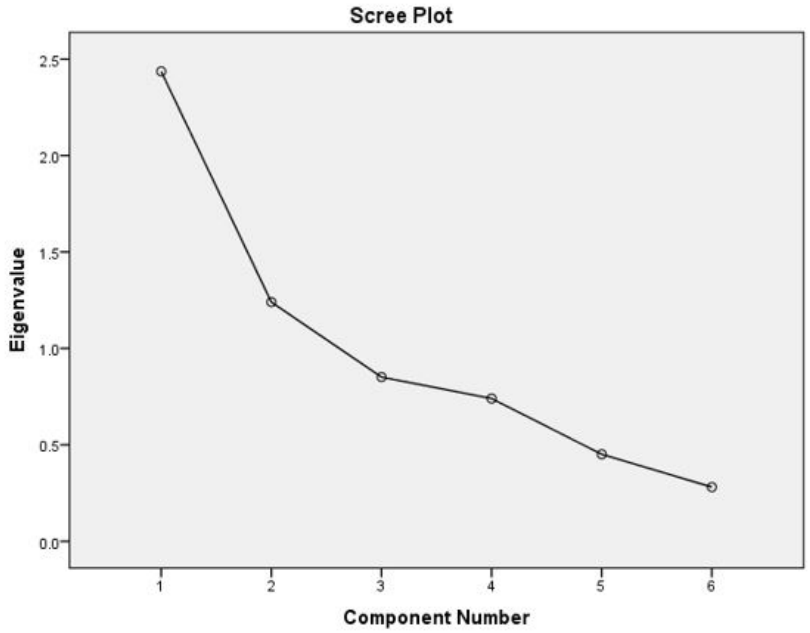


**Figure 4.4** A scree plot of the factor analysis shown in Fig. 4.3, showing the magnitude of each eigenvalue (or component).

Component Matrix

| | Component | |
|---|---|---|
| | 1 | 2 |
| Quality | .843 | -.011 |
| Width | -.018 | .777 |
| Reverb | -.126 | .685 |
| Real | .863 | .161 |
| QualityPlace | - 587 | .318 |
| Mimic | .787 | .200 |

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

**Figure 4.5** This shows the correlation between each test items and the components found in the factor analysis of the test described in Appendix B.

highly correlated to that component. For example, the factor analysis showed that quality, quality place, realism of the sound, and the ability to mimic a live recording all highly contribute to the first factor. Further, ASW and perceived reverberation contribute to the second factor. From this analysis, the researcher may consider that these are two separate areas that the test is measuring.

Before using this factor analysis, the Cronbach's alpha for the test was 0.56, which signifies the test was not highly reliable. However, additional Cronbach's alphas were measured comparing the items in the two factors; quality, perceived realness of the sound, and the ability to mimic the live recording returned an alpha score of 0.82, which is almost ideal. Similarly, ASW and reverberation also reported a much improved alpha score. By computing the two different components found in the factor analysis, this test was found to be highly reliable. A similar process can be used in any other test evaluation.

### 4.2.2 Validity

There are multiple aspects to validity that have to be ensured. One of these is called content validity, which is a measure of how much the test covers all aspects of the given hypothetical construct being tested. So if the operational definition of quality previously given seems insufficent, it can be said to not have content validity. Also, if there was a question such as "What is your favorite color?", it would have no content validity when evaluating how a loudspeaker is colored because it would not measure the loudspeaker's properties.

A measure to test each question's content validity is called a content validity ratio (CVR). It is a relatively simple process to implement. To do so, one gathers a group of experts (for loudspeaker quality, imagine a group of audio engineers). The more experts the better, but a group of five is sufficent. They rate each question as being either essential, useful, or not necessary. Following their rating, the total number of essential votes for each question are tallied and use the following equation provides the CVR score:

$$CVR = \frac{E - \frac{N}{2}}{\frac{N}{2}} \tag{4.2}$$

where E is the number of votes for essential and $N$ is the number of participants. This returns a number from -1 to 1. The required minimal CVR score for high content validity has been researched [6]. A table that matches number of required votes for essential relative to the number of participants is in appendix C. These numbers are less important than a good intuition of what the experts have said. If two questions receive high CVR scores that are almost identical, only one should be used. In addition, a question may receive many useful ratings but few essentials. Despite the low corresponding CVR score this question would receive, it should still be considered[3]. Every question may have high CVR scores,

[3]To the author's knowledge, there is no method or equation used that considers the number of votes for useful.

but it is important to consult the experts carefully to see if there is a need for additional questions to give a complete definition of the hypothetical construct.

Another aspect is criterion-related validity. This is a measure of how well the results of the test relate to a known criterion. For measuring loudspeaker quality, imagine a highly respected audio engineer that is an expert on hearing loudspeakers quality. His or her judgement can be the criterion. The test results should then say loudspeakers have the same relative quality to each other as this expert says they do. Another possible criterion may be a respected and proved test that provides similar information as the newly made test. If the new test agrees with the proven test, it has high criterion-related validity. Every criterion can have issues however; it is nonetheless crucial to choose at least one criterion to test the validity.

To measure the criterion's relation to results of the test it is useful to perform statistical analysis on its results. Chapter 3 covers useful statistical procedures. Correlation values are useful to validate the criterion and the test's results. If the criterion and new test are supposed to measure the same thing, they should have a high correlation. Factor analysis can help evaluate this as well. These statistics can help give hard data on the success of the test.

One final aspect is construct validity. It is the overarching question of validity: does the test correctly measure what it is supposed to measure. In order to have construct validity, the test must have both content and criterion-related validity. While measuring a test's content and criterion-related validity can begin to assure the construct validity of a test, it requires a bit of intuition to fully know if a test truly has construct validity beyond the numbers used to represent it. Every test will to some degree fail to completely measure what it tries to, and whoever is using the test will need to understand its limitations. The statistical analysis gives some helpful information as to these limitations, but statistics are limited. In

the end, the best tester and researcher will evaluate the results of the research themselves. For example, if the objective measure of A-weighting might inappropriately represent loud low-frequency content, and the researcher could discover this by comparing A-weighting levels to what he or she heard. Similarly, with subjective measures, the researcher will need an intuition beyond just a set of known protocols and methods. This intution comes mostly through experience, and by carefully listening to the different sources being compared.

## 4.3 Implementing the Test

Once a test is made, its implementation is crucial. A perfect test will have unreliable results if its implementation is incorrect. This is similar to testing objective quantities. Even if a researcher uses a top-of-the-line signal analyzer to measure some objective quantity, if the wires, transducers, and testing environment are not set up correctly, the data gathered by that signal analyzer is useless.

Another consideration is how human participants color the signal received. Those in technical fields will generally treat their measurement systems (microphones, signal analyzers, and so forth) as something that has an input, a linear impact on that sound, and then a simple output, as shown in Fig. 4.6(a). The technical background seems to cross over into how human participants are treated; one may simply change out a human for their measurement system [Fig. 4.6(b)]. In reality though, there are many other input variables distinct from the experimental input; each person himself or herself colors the signal differently [Fig. 4.6(c)]. In fact, human participants may be vastly different. To relate this to an objective acoustical measurement, it would be like attempting to measure some signal using a different low-cost microphone brand and model for each recording while there is sporadically noisy construction going on in the background.

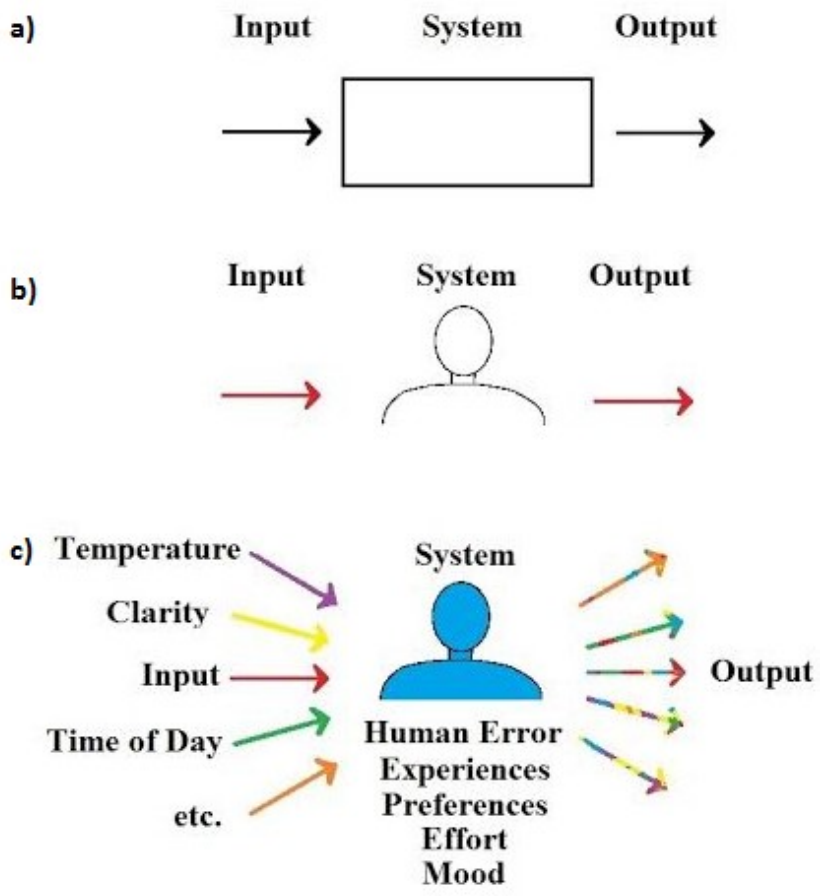


**Figure 4.6** (a) A typical linear system with some input, a process done by some system, and an output. (b) A common mistake is to treat human subjects as single-input, single-output linear systems, where they will simply take an input and generated an uncolored output. (c) A better illustration of how humans alter any input includes many other inputs, considerations of the individuals, and a wide array of outputs.

This section will outline several different ways to make the implementation of tests more reliable and repeatable. Humans are often "noisy systems" as compared to scientific measurement equipment and many things can cause potential errors in the measurements. There has already been a large amount of research investigating which aspects of the environment or individual can cause problems, and how to properly limit their potential impact.

### 4.3.1 Minimizing Extra Variables

**Radomization** One of the simplest ways to limit the impact of outside factors in the data is proper randomization. For example, if participants are told to listen to loudspeaker A, rate it, and then listen to a loudspeaker B and rate it, the participant's assessment might be partially dependent on the order in which the loudspeakers are presented. To limit the impact of order it is a good idea to have half the participants listen to loudspeaker A first and loudspeaker B second while the other half of participants listen to loudspeaker B first and loudspeaker A second.

If there are more than two loudspeakers, if possible use all iterations. So if there is a loudspeaker C, use each of the six possible iterations equally. Upon the introduction of a fourth and fifth element however, there might be more possible iterations than the research is capable of executing. To optimize the order when it is not possible to use each iteration there are multiple techniques. One of these is a balanced Latin square. To use it, follow this procedure:

1. Give a number to each possible different condition (i.e. loudspeaker A is 1, loudspeaker B is 2, etc.) and let $n$ be the total number of conditions.

2. Make an empty n x n table.

3. If n is even,

(a) The first row is 1, 2, n, 3, *n* - 1, 4, *n* - 2, . . .

(b) The second row is found by taking the value from the previous row and adding 1. If the new value is equal to n + 1, instead of its value becoming n + 1 it becomes 1.

(c) Repeat for each row.

4. If n is odd:

(a) Create two empty n x n tables.

(b) Follow step 3 to fill the first table.

(c) For the second table, use the process in step 3 but with each row being reversed in order.

(d) When administering the test, use all of the orders in the first table, and then use the orders in the second table.

Another potential method to control the randomization is testing blocks, or specific sets of orderings to test in. For instance if there are five loudspeakers labeled A, B, C, D, and E being sequentially tested, the order they are tested to each participant can be in one of three blocks:

1. A B C D E

2. E D C B A

3. B D A C E

These blocks would have blocks that change which loudspeaker is presented first, and each loudspeaker is measured in-between different loudspeakers. As there are only three iterations, each iteration may be tested multiple times. The advantage to this method is

that the impact of the order can be measured statistically as there are sufficient number of iterations to test between.

**Standards** In subjective evaluations of loudspeakers, the term standard is often misleading. Many standards are often used as general guidelines in order to perform research, but the data gathered in a subjective test may be used even if every guideline is not followed perfectly. This is explicitly stated in AES20 [7], and can be assumed for most standards. The standards with rigid guidelines are generally in medical situations (i.e., hearing aids and hearing tests).

While the standards may not be strictly enforced, they serve as an excellent starting point for research development. A list of various standards and their corresponding applications are given in Appendix D, which also contains a comparison chart. Generally though, the best way to find any standard that might be needed in any given test is to find previous research that is similar, and determine the standards used (if any).

**Room Design** It is obvious that the design of a room will alter the acoustics of any sound output in it. Any loudspeaker would sound different in a small bedroom than in a large concert hall. Because of this, there is an abundance of research on the impact of the room on subjective evaluations [2] [8] [9] [10] [11]. However, it has been found that humans are very capable of adapting to a room and judging the acoustical source (such as a loudspeaker) without the room largely influencing their judgement [see *Sound Reproduction* chapter 11 [2]]. The main considerations are that the room does not eliminate part of the loudspeaker signal. This applies in a few ways. First, the background noise cannot be so high that it masks some of the subtle qualities of any of the loudspeakers. If it is not possible to hear those aspects of the loudspeaker, it is not possible to judge those aspects. Further, it is important to set up the room so that the impact from low-frequency modes are limited at

the listener's position. Factors from low frequencies makes up about 30% of the overall rating of a loudspeaker's performance [12], and so if there are a large number of nodes or antinodes at the listener's position the true quality of the loudspeaker may be hidden. Further, there should not be large extremes in any characteristic of the listening room (i.e., reverberation, width, length, height, shape, etc.). Some reverberation in a room is crucial to a loudspeaker's sound so an anechoic chamber would be a poor location. On the other hand, too much reverberation can mask acoustic quality so it is inappropriate to use a reverberation chamber for subjective evaluations. Similarly, a room should not be too small or too large. Participants should be at least three wavelengths away from scattering surfaces [13], but early reflections arriving too late because of distant walls will cause issues as well. Essentially, the room should not have any extreme influence on the acoustics at the listener's position.

If specific research necessities create different concerns, a researcher may adjust for them. If each loudspeaker (or other acoustic source) is being judged in the same room, the room should not significantly influence participants' ratings of the loudspeakers (as long as the previous requirements have been satisfied). Human hearing is capable of adapting to its environment, so the impact of a room is limited. For peace of mind when considering room design, Appendix D has a list of standards for different types of tests, which outline *suggestions* of an "ideal" testing room. The word suggestion is used because these standards do not apply to every possible scenario, and they may be altered according to the researcher's needs. Data may still valid be if the room does not perfectly fit one standard.

**Loudspeaker Position** The position and orientation of a speaker or any acoustic source may change sound perceived by a listener. If multiple acoustic sources are being assessed simultaneously, the location of each loudspeaker should rotate between each location a loudspeaker is being tested (see the randomization section). For instance, if two loud-

speakers are being tested, the loudspeaker on the right and the loudspeaker on the left should switch places. The listener should be located in the direct acoustic far field if possible. For most frequencies, this generally requires a distance of about two meters from the source when the room is also considered, but each test will be slightly different. Unless the off-axis response is being evaluated, the loudspeakers should face the listener, much as they would if the listener were not rating them.

**Double-Blind Test Setup** Blind here means removing knowledge that could bias a participant or test administrator. So if loudspeaker quality is being tested, blind means not knowing which brand and type of loudspeaker is being tested and not seeing the loudspeakers' appearance. Double blind means that both the participant and the person administering the test have this ignorance of the potentially biasing information. A simple way to do this for the participant is to use an acoustically transparent (or in practice, translucent) but visually opaque screen so that listeners cannot see what they are evaluating. As for the test administrator, he or she generally should read from a script when giving instructions. To further limit bias, the test should either be computer automated or not need the test administrator's assistance once the administrator knows information that might bias the experiment. For example, the administrator should not know the order of the loudspeakers being evaluated when explaining the procedures to the listener. If the administrator must be involved in arranging the loudspeakers, the instructions should have already been given before the test begins.

Double-blind setup is crucial when evaluating acoustic sources, especially loudspeakers. Research done by Floyd Toole and Sean Olive [14] provides an in-depth analysis of the ability of listeners to be objective in their evaluations of different loudspeakers. One set of tests had loudspeakers placed behind an acoustically transparent screen. Another test was done where the listeners could see the loudspeakers. It is pertinent that these were profes-

sionally trained listeners, taught how to listen to loudspeakers and evaluate them without bias. They were experts in the audio field. Despite this fact, their knowledge of what the quality of each loudspeaker "should" be according to its brand/price drastically influenced their judgement. Without an acoustically transparent screen, the listeners reported the more expensive loudspeakers sounded better. The more aesthetic loudspeaker enclosure also had better reported sound quality. Professionally trained listeners were unable to negate what they saw, so no participant will be able to escape this. This also means any test that does not blind the participant and does not consider the impact of participant bias cannot be trusted.

Blinding the administrator is also important. If the administrator has a bias towards one loudspeaker, it is possible that this bias will be shown in something as simple as his or her tone when describing the protocol to the participant. This in itself could bias the results towards what the administrator's bias, and so generally data gained from research that is not blind to the administrator, as well as the participant, is suggestible.

**Listener Qualifications** The participant evaluating the acoustic source is one of the most likely sources of errors in the data. As a result, there are many considerations that should be taken into account. First, they should have minimal or no hearing loss. If someone is incapable of hearing certain frequencies, then their response could be drastically different than that of the average listener. No amount of test setup or listener preparation can compensate for a participant not being able to hear something. Floyd Toole put it strongly saying, "In fact, anyone with gray hair, especially if they are in the professional audio business, should be considered suspect as an arbiter" [2]. Simply put, unless hearing loss is part of the study, do not use individuals with hearing loss.

It is also helpful if participants are accustomed to what they are listening for. Again, the example will be evaluating loudspeakers. Someone not accustomed to critically listening to loudspeakers will not be nearly as good at evaluating a loudspeaker as an audiophile. It

has been found that the most reliable and observant participants were those who identified as hi-fi enthusiasts, the next best group were musicians, and then the worst were those with no hi-fi interests or musical training [15]. Further, those who work with audio files such as recording or mastering engineers are often better than hi-fi enthusiasts or audiophiles who passively listen.

There is one way to cause participants to become much more reliable, and that is to train them how to critically listen. There is exstensive research into training programs that shows them to be very beneficial [12] [16] [17]. One researcher found that the number of trained listeners needed to produce a certain confidence interval in the data can be less than the number of non-trained listeners needed by a factor of seven, and that training may still be effective up to one year after completed [16]. Trained listeners come to the same conclusion as untrained listeners when comparing quality of loudspeakers, but routinely give lower quality scores [2].

Listener training follows a very simple process. A listener will listen to an unaltered audio track and compare it to the same audio track that has been altered. He or she will then report in some fashion what the alteration is. Sean Olive [12] designed a training program wherein a track would be altered by four different filters that were either a boost or cut to low, mid, or high frequencies to alter the track. This program was recreated by the author at BYU (see Appendix A). Listeners who used it found a noticeable difference in their ability to hear subtle differences in different sounds. Some just noticed they could better hear changes in quality when listening on their MP3 player as compared to better audio files. Some people who worked as sound engineers for BYU were better able to equalize sound for optimal quality.

Creating a listener training program can be difficult. Fortunately, there is a free listening training program that can be downloaded online (harmanhowtolisten.blogspot.com). It is

distributed by Harman International and is called "How to Listen."

**Program Selection** When evaluating something like a loudspeaker, one must determine how well it is able to reproduce any acoustic signal. However, it is not possible to use every possible acoustic signal. It is consequently necessary to use an audio selection that is able to represent broad characteristics of the quality of the loudspeaker.

Generally, there are two main types of sounds that apply to human audiences: speech and music. For speech, many sound tracks already exist. For general reading passages that are phonetically balanced, the Rainbow Passage and Marvin Williams passage are often used [18]. For evaluations of hearing speech in noisy environments the hearing in noise test (HINT) is very effective [19]. (The HINT is often used to evaluate hearing aids.) There are also audio tracks that exist which contain made up words that sound similar to English, but have no meaning. These files are all useful to analyze a loudspeaker's clarity. Another type of speech test will have a word reproduced, and will ask the participant what word was said from a list of words that include the reproduced word and five other words which rhyme with it.

Music selections are less specific. There is no predefined program to use for most types of evaluations. Overall, pop or light rock music is excellent. Classical music seems to lack compared to pop [12]. Too hard of rock or electronic music may have too much noise for a listener to hear how sound quality is changed. The selected song should fulfill a few requirements. It must be mixed and mastered extremely well. As discussed earlier, things such as ASW contribute to sound quality and this can only be heard if the music was mixed correctly. There also should be a minimal amount of clipping in the file. It is common practice now to amplify music until it is too loud and may thus have amplitude clipping. The file should be inspected to see if this has been done extensively (this can be done by using audio software such as *Audacity* and simply zooming in on the wave

form of the audio signal for the loudest passages). The music should also have a full set of instruments, meaning percussion, bass, guitar, and voice as a minimum, or some sort of equivalent to each. Music with fewer instruments will not require the loudspeaker to use its full bandwidth or fully expose its fidelity. An excellent example of all these considerations is "Hotel California" by The Eagles in the album *Hell Freezes Over*.

Another important consideration is a participant's familiarity with the music. In past research [12] it was found that many songs by Tracy Chapman were excellent to use for loudspeaker quality assessment. However, some of these were the most difficult selections for listeners at BYU going through a listening training program. Their ability to correctly recognize the alterations to the music were significantly less than the other selections. This was also true of the THX sound clip played before movies in theaters. Although many sound engineers claim that it is an ideal track to hear sound quality, the listeners being trained seemed to have a hard time hearing alterations to it. It is the author's opinion that this is due to the lack of participants' familiarity with how the Tracy Chapman music and THX sound clip should sound. Almost no listener normally listened to Tracy Chapman or even knew who she was, and most did not previously use the THX sound to evaluate sound quality. They did better though on songs similar to Tracy Chapman's music that were more modern. So music also should be selected that the listener is used to. To aid in this it is helpful to give the listener copies of the audio selection in the evaluation so he or she can listen to them beforehand and become accustomed to how they sound.

On a final note, participants do not have to enjoy the music selected. In the training done at BYU, many Taylor Swift songs were excellent at helping listeners hear changes in the color of a sound. Not everyone enjoys Taylor Swift's music, but most have heard it. Listeners "can detect flaws in the reproduction of music in which . . . [they] find no pleasure" [2].

**Electroacoustic Considerations: Level, Equalization, and Number of Channels** How the sound signals are set up is just as important as any other item listed. For the loudness level of the loudspeaker, it should not be so loud as to hurt the listener but plenty loud that every aspect of the loudspeaker can be heard. This is quite simple; imagine listening to a song through a MP3 player while it is quiet. Likely some of the low frequencies, high frequencies, and background instruments are so quiet they cannot be heard. When the volume is turned up, though, these aspects that were not heard before will come out in the music. It is the same for loudspeakers. Because of this it is crucial that participants all hear the loudspeakers at the same loudness level. If not they will literally hear different sounds produced by the loudspeaker. Therefore, a high-sensitivity loudspeaker would need to have its input voltage reduced if being compared to a low-sensitivity loudspeaker until the two sound outputs are the same [2]. Their sound pressure levels can be measured, and generally, A-weighted levels better predict how loud the sound will be to the participant. There are better psychoacoustic measurements of the sound level heard by an individual, but A-weighting is sufficient because it is designed to match subjective listening.

The spectral content of a loudspeaker generally should not be equalized. If the spectral content is equalized, it must be understood that the equalization then becomes a potential variable in the experiment. Furthermore, the equalization limits the generalizability as it may be very different than the equalization that would be used by an end user.

When measuring quality generally using mono is better than stereo or multichannel setups. Largely, winners of mono tests win stereo tests [2]. When evaluating sound envelopment considerations such as ASW though, stereo might be necessary to create the impression of a real sound stage. The researcher needs to choose if the color or the spaciousness of the sound is more important. The test may also use mono to measure color and stereo to measure envelopment if the two are not simultaneously measured.

**Test Administration** When establishing the logistics of how the test will be administered, whatever is decided upon needs to become very repeatable. Just as all of the questions must have words clearly defined for every participant, the procedures must be explained in the same way to every participant. If the procedure is somehow changed or explained differently, that can influence the results in the test in an unpredictable way.

To help with this, make sure that the following things are already established sufficiently to be highly repeatable before administering the test: (1) recruiting process, (2) the location where the participant will go to perform the test, (3) how they will be introduced, (4) a script of how the project will be described, (5) what tasks the test administrator will and will not do as the participant performs the test, and (6) how the participant is to be debriefed. In these considerations, the procedure does not have to be read word for word from a script, but if it is not, the administrator will need to have practiced the process of describing the research so that it is described the same every time.

Once the procedure is decided upon, a pilot test should be carried out. No matter how much preparation there is beforehand, a pilot test will reveal something that needs to be altered. Once the actual testing has begun, if there are changes to the procedure, the data from before the changes and after the changes generally should not be combined into one total set of data. A pilot test helps eliminate this issue.

**Number of Participants** In order to generalize the results of the research, there needs to be a sufficient number of participants. The required number does vary based on test design and, for loudspeaker tests, the skill of the listeners. When pooling from the general population, there should be a minimum of approximately 20 participants.

### 4.3.2 Consdierations for the Participant

**Participant's Comfort** The test needs to be set up such that the participant is able to give an unaltered opinion. One important consideration with this is that the participant needs to be familiar with his or her surroundings. Therefore, a participant should have some time in the loudspeaker evaluation room to adjust to its environment. (If the description of the protocol is given in the room, that is sufficient.) They should also be familiar with the protocol before they are assessed. This might mean giving them a sample question as practice that will help them feel confident in the procedure. If this is done, it is generally inappropriate to use the response to the practice question in the data analysis. In addition, as mentioned in program selection, participants should be accustomed to what they are listening to.

The room should be comfortable enough that the participant does not have to think about the surrounding. The room should not be too hot or too cold, but optimized so that the participant does not notice the temperature. The chair that the participant sits in should be comfortable, but not so comfortable that it will put him or her to sleep. No condition should be so extreme as to steal away from the participant's attention.

**Limit Participant's Time Involvement** The participant will only be able to give good results for a limited amount of time. If they are required to focus on sounds, think, or speak for too prolonged a time, they may become fatigued. Accordingly, participants should generally spend less than one hour in subjective assessments. This includes time spent on any practice rounds. This is not a rigid rule, but is an important consideration.

### 4.3.3 Take Advanatage of Extra Variables

**Measure the Extra Variable** It is hard to know which extra variable may influence the data. While it is good research technique to attempt to remove extra variables from the independent and dependent variable(s), inevitably there will be some additional variables in the research (such as gender, ethnicity, or age of the participant). Because of this, much consideration should be given to ascertain extra variables in the research and take note of them. For instance, even after randomization, the order of presentation may have an impact. All such variable have the potential to color the data and, if feasible, should be measured.

This will often include creating a survey of some sort to help gather information about the participants that might impact their judgement. Gender, ethnicity, and age can be determined, as well as musical training, preferences in music, or their general well-being the day of the test. One rule-of-thumb of good subjective research is this: if there is an opportunity to gain more data that may be significant, take it. It is hard to know what will end up being part of a key discovery.

**Measuring the Impact of Extra Variables** While it may be hard prior to assessment to know how any extra variable will impact the data, it is possible in the post processing of the data to see if there was an extra variable impact and what it was. The process is relatively simple. Chapter 3.2 covers how to compare means using things such as $t$ tests, ANOVA's, or post-hoc analysis. One goal is to see an impact from some change in an independent variable, and so it hopes to have a $p$-value of 0.05 or less. Another goal may be that some variable does not influence the data. One simply performs the same statistical process and if there is a large $p$-value, then that variable did not influence the data sufficiently that it can be measured. The $p$-value should be much more than 0.1 before it is reasonable to neglect its influence. It is important to note, though, that this does not imply that this extra

variable would not impact the results in other iterations of the experiment, only that there was no significant difference detected in that set of data. If the goal is to definitively say that something is not a factor, other techniques must be used. Another consideration is that if a small number of iterations are tested it is also possible that a type II error is committed. This means that a high $p$-value would not show that reseach was not impacted.

One major advantage of checking to see if extra variables had an impact in the measurement is that it can be a win-win situation. If there was no measured impact from that variable then the researcher can be more confident in using the data.[4] If there is an impact, that is something interesting that may be appropriate to report. For example, if gender plays a substantial role in the preference between loudspeaker A and loudspeaker B, the marketing of that loudspeaker would be vastly different. Alternatively, if the measurement is used for research, that variable might be something that can be further researched and possibly published. Regardless, checking the extra variables provides useful information. The major limitation is usually the time commitment of the researcher.

## 4.4 Conclusion

When designing, evaluating, and implementing a test for subjective evaluation, there are many considerations that differ from objective evaluations. When these differences are considered, very effective tests can still be made. The guidelines in this chapter do not cover every aspect of test design and implementation, but it is sufficient as an introduction.

[4] It is important to remember that if there is a small number of iterations to test over, a type-II error is possible.

# Chapter 5

# Conclusions

While there are many intricacies involved in performing subjective evaluations that most with technical backgrounds are not aware of, it is possible for them to learn and implement these techniques. This thesis provided an overview so that a novice may begin to perform subjective evaluations of acoustic sources. By learning these techniques, he or she can incorporate subjective opinions into research with human particpants that are ethically treated and with data that are correctly gathered and analyzed.

It is of primary importance to understand the ethical considerations of working with human participants. To safeguard participants' safety, it is necessary to gain approval from the IRB board. This has a specific purpose, an understanding of how they evaluate research merits aids and accelerates the approval of research proposals submitted to them.

Statistical analysis must also be done conducted when considering the responses of participants. Often, participants' responses are compared using measures of central tendency such as mean values. When means are compared, it is also crucial to consider measures of spread to see how exclusive each mean is. A large series of tests (such as *t*-tests, ANOVA, and post-hoc analysis) are able to demonstrate the amount of confidence (through a *p*-value) that each mean is statistically different from another mean value. Correlation between vari-

ables may also be measured by using Pearson's *r*.

With an understanding these items, it is important to properly design, evaluate, and implement subjective tests. When developing a test of a subjective measure, an operational definition often has to be given to the hypothetical construct. The operational definition has to then be divided into appropriate questions that will meaningfully illuminate it. Once a test is made, its reliability and validity must be tested. The validity's content must be relevant to the operational definition being tested and should cover the full definition. A test's reliability can be tested by using Cronbach's alpha and factor analysis. The validity can be further tested using some known criterion as a reference point. Once a test has been decided upon and its reliability and validity are ensured, its implementation is crucial. Through careful planning the test can be implemented in such a way as to limit the impact of extraneous variables.

This thesis serves only as a brief introduction into the world of subjective evaluations. There are vast resources beyond it if a researcher wishes to become truly proficient in subjective testing. The thesis covers only the most basic of statistical procedures and testing methods an is intended for those with a technical background in acoustics. It provides enough information to help prevent grave errors in test design and provides tools so that an initial analysis of the data can be performed and reported.

Beyond this, researchers should strive to read more books in testing methods and statistical analysis. Some helpful books are *Sound Reproduction: the Acoustics and Psychoacoustics of Loudspeakers and Rooms* [2], *Perceptual Audio Evaluation: Theory, Method, and Application* [20], and *Multivariate Analysis for the Biobehavioral and Social Sciences* [5]. There is a wide array of techniques that, once learned, enable a researcher to becom efficient and highly effective. While this thesis serves as a foundation, further understanding of the presented concepts can greatly improve research efforts.

The techniques presented in the thesis are simple. However, by using them proficentlyy, a new area of research can open up to researchers in technical fields. They have been able to greatly enhance the author's research, and can help enhance the projects of others who read this material and apply it judiciously.

# Appendix A

# Listener Training Progam

As part of this research effort, the author developed a listener training program to be used at BYU. It was largely based on a similar training program developed by Sean Olive [12]. Its goal was to enable listeners to hear subtle differences in similar sound files in preperation for their participation in the research described in Appendix B.

The training process was very simple. A group of songs were given a 3 dB boost or 3 dB cut in a limited band with a $Q$-factor of 0.66 at 500 Hz or 2 kHz, or a broad shelve 3 dB boost or cut at 100 Hz or 5 kHz. Those being trained would use a program created by the author to click a button to hear an unaltered version of the music. Beneath the button were four buttons each with a random boost or cut from the options above. There were also four graphs that showed the four filters that had been used. The layout is shown in Fig. A.1. The buttons would initially be gray and the trainees would select which button's music corresponded to which filter on the left.

After trainees felt they had the correct connections between the sound files and the filters on the left, they would submit their answers. A screen would then appear and tell them which of their guesses were correct and which were incorrect. If they did not get every button linked to the correct filter, the screen would show a tip and ask them to try

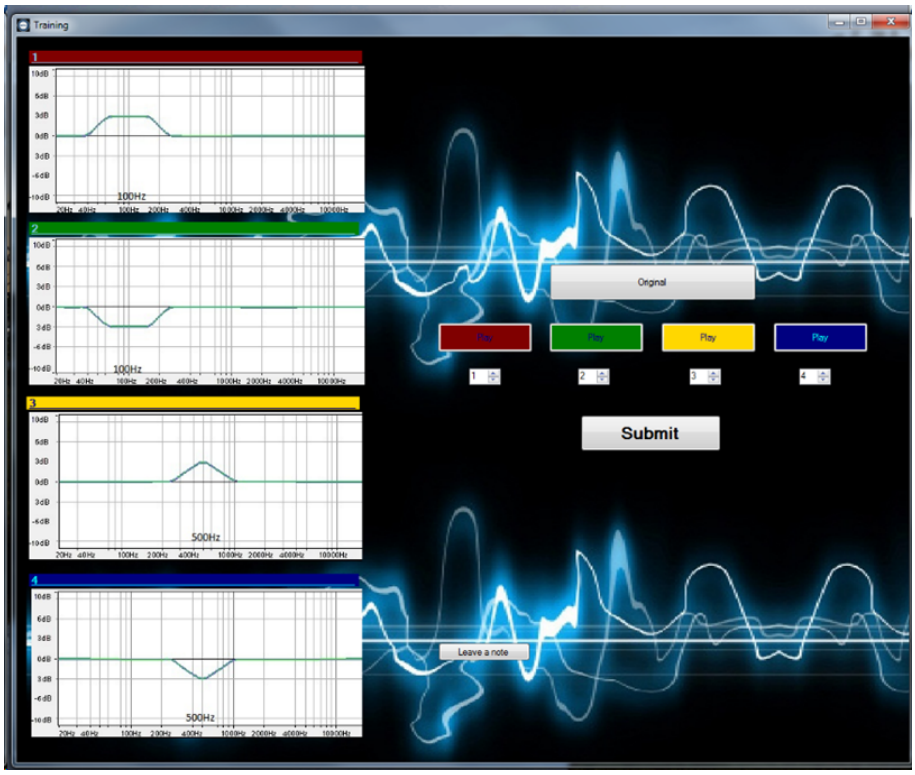


**Figure A.1** The page trainees would see in the listener training program developed at BYU. Their job was to match what they heard when they pressed the bottom four buttons to the four filters illustrated on the left side.

again. This continued until the trainee was able to get all the answers correctly.

As the trainees at BYU were being prepared to perform a subjetive evaluation, they followed a specific regimen. They were required to perform eight training sessions, with each session containing eight training pages as shown in Fig.A.1. Each session would consist of a song being used twice. The first time a song was used, its filters were randomally chosen (see randomization section in section 4.3.1). The same song would then be used on the next page with the filters that had not been used in the first iteration. Another song would then be used with the same process. In total, four songs were used each day.

The first set of training was done with all trainees together to ensure they properly understood the process. Each training session following the introduction, trainees used


**Figure A.2** The opening page trainees for the listening training program developed at BYU. Trainees could access their training from this page.

an opening page, shown in Fig. A.2. Participants were randomally assigned to begin on different sessions, with the exception that none started with session one as it was used for the introductory session. Participants were required to do perform two training sessions within a week, but were not allowed to do more than one training session within a 48-hour time period. Each training session lasted 10-30 min., with the time depending largely on the past mixing and recording experience of the trainee.

The songs used for the experiment were the following:[1]

1. Baby by Justin Bieber
2. Stars and Stripes (Arrangement by Cy Pane, from Firstcom Music)
3. West of Hollywood by Steely Dan
4. Secrets by One Republic - Instrumental (covered by Piano Guys)

[1]The song selection was chosen to represent many types of music. This list is not meant to suggest which music should be used, but merely to report what the author used.

5. Beethoven's 5th (performed by New York Philharmonic directed by Leonard Bernsteing in 1961)
6. Red Camaro by Rascal Flatts
7. White Horse by Taylor Swift
8. Imperial March by John Williams
9. Black Friday by Steely Dan
10. Somebody to Love by Justin Bieber
11. The Remedy by Jason Mraz
12. Jurassic Park Theme by John Williams
13. You Belong With Me by Taylor Swift
14. Cousin Dupree by Steely Dan
15. Superman Theme by John Williams
16. Fast Car by Tracy Chapman
17. Point of Know Return by Kansas
18. Carry on Wayward Son by Kansas
19. Radioactive by Imagine Dragons
20. Why by Tracy Chapman
21. With a Little Help from my Friends by the Beatles
22. Cougar Fight Song played by BYU marching band
23. On Top of the World by Imagine Dragons
24. Pink Noise
25. Selection from BYU's wind symphony
26. THX sound
27. The Wall by Kansas
28. Secrets by One Republic (covered by Piano Guys with voice of Tiffany Alvord)
29. Love Story by Taylor Swift
30. Cougar Fight Song played by BYU marching band
31. Duel of the Fates by John Williams

Items 29 to 31 were the items used for the introductory training session, with Love

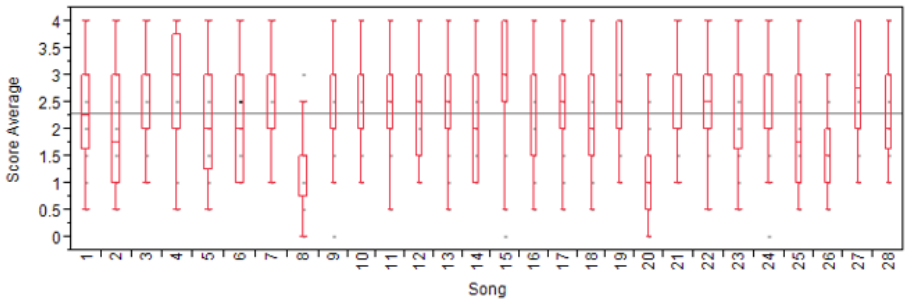


**Figure A.3** The initial score of trainees on their first attempt to correctly match the altered music with the corresponding filter, with each correct answer receiving one point.

Story by Taylor Swift being used four times to aid the initial understanding.

The initial score of trainees on their first attempt to correctly match the altered music with the corresponding filter, with each correct answer receiving one point, were recorded. The mean value for eah song is shown in Fig. A.3. The selections that the participants seemed to have the most difficulty with were 8, 20, and 26. "THX" sound. There would likely be more drastic effects between each mean value, but as trainees were improving over time in their ability to correctly guess the corresponding filters, the corresponding data was quite inconsistent. It is worth noting that the songs with the most issues were either uncommon to a college-age group or instrumental pieces. In addition, the average number of attempts to correctly match the sound clips to the corresponding filter of the first and last training session were compared. Participants were found to improve by an average of 0.29 attempts ($\sigma = 0.13, p = 0.04$).

# Appendix B

# Example Evaluation of a Subjective Source

Over the course of the past few years, the BYU acoustics research group has developed a new method to measure the directivities of musical instruments. The new method generates a frequency dependent directivity balloon containing 2522 data points. There was interest at BYU to see the impact of using this new directivity in archeticual acoustics modeling software such as EASE. The author thus performed a subjective evaluation of the approach.

To do this, a binaural recording was taken of a trombonist, using a KEMAR mannequin in the De Jong Concert Hall at BYU. He performed four different pieces multiple times at multiple dynamic levels. The trombonist was given a metrenome that also showed his pitch so that he was able to keep uniform timing and pitch for every iteration. The same process was then repeated in an anechoich chamber. The anechoic recordings were subsequently convolved with two binaural impulse responses of a model of the De Jong Concert Hall generated by EASE. One convolution assumed an omnidirectional source and the other used the new directivity data measured at BYU. The different versions of the recordings


**Figure B.1** This is the opening screen listeners saw while performing the test.

were then compared to find the segments wherein the musician was most similar between the De Jong recording and the anechoic recording. The most similar sections were cut into smaller sections to be used for the later evaluation. Background noise from the De Jong recording was also isolated and added to the anechoic recordings and every recording was normalized to have the same RMS amplitude.

A test was then designed to asses the perceived quality of each recording and the realism of the sound. The test was evaluated by trained listeners (see Appendix A). It was administered using Sennheiser HD 650 headphones and participants answered questions while in a small anechoic chamber to limit background noise. To decrease electronic noise, an outboard digital audio interface (Lexicon Omega) also used.

The test was administered through a program developed using C#. The opening page each listener saw is shown in Fig. B.1. He of she entered an ID number that the program would use so that the order of all files presented was properly randomized. The test was designed so that each button would only become usable once the previous section of the test had been completed. Listeners first provided some background information about themselves. The form they filled out is shown in Fig. B.2.

After the listeners filled out the questionnaire , the author as the test administrator ex-

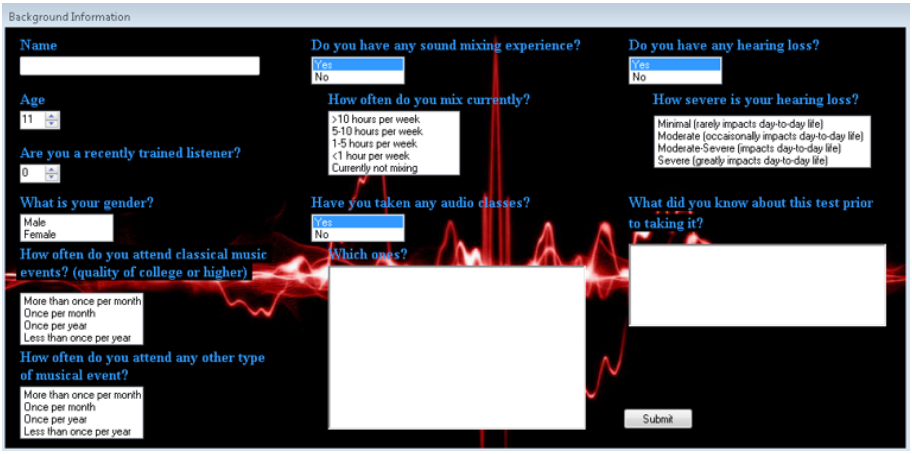


**Figure B.2** The questionnaire filled out by listeners to get some background information.

plained the procedure for the first half of the test. The explanation was given with a prepared PowerPoint presentation so that the explanation given to each particiapnt was the same. They were shown the evaluation screen (Fig. B.3), and then taught what was meant by each question. In case they had questions following the directions, buttons with question marks would open up a new window to give the same explanations previously given. Listeners were asked to assess the overall quality on a 0-10 scale, the comparitive ASW of the three files, the perceived reverberance using a Likert scale, and how real the file sounded on a 0-10 scale. The quality scores were each required to be different (so the listeners could not give the first version the same score as the second and third version, and so forth). The three different versions of the recording were randomally linked to one of the three buttons. After completion of one evaluation, the listener then repeated the procedure three more times with different selections recorded from the trombonist.

After completing the first four evaluations, the listener was given a new set of instructions for another test. He or she was informed that one recording was a true recording in the De Jong concert hall and that the other two recordings were made using room modeling software. They were then shown a new test form (Fig. B.4). This part of the test asked the

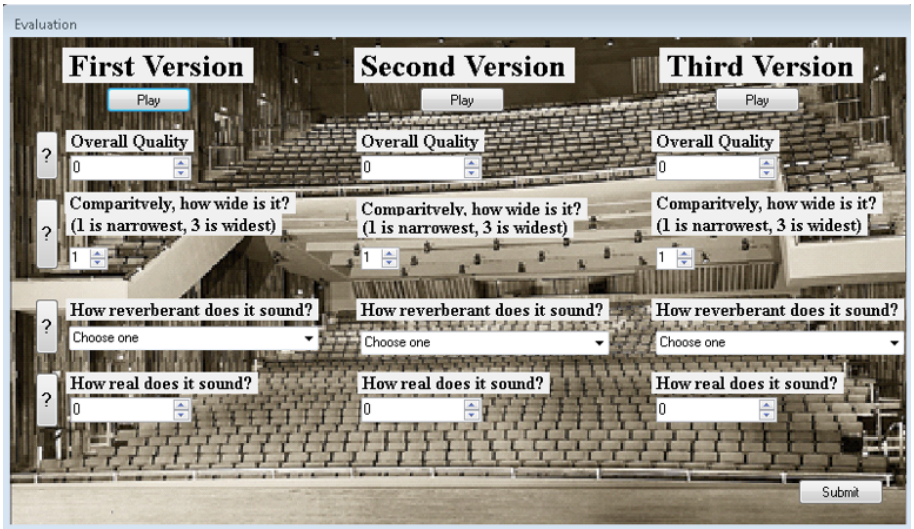


**Figure B.3** The first part of the test the participants performed.

listeners to only compare the two recordings convolved with the model of the De Jong, but it gave them the option to hear what the live recording of the De Jong Concert Hall sounded like. They were then asked how well the recording mimiced the De Jong concert hall on a 0-10 scale. If they gave equal scores to each recording they were asked to indicate which one seemed to mimic better. Following this set of instructions, they then performed four tests using this form on the same four selections used on the first half of the test.

Following the tests, the data were gathered and analyzed. First, the mean values were looked at (Fig. B.5).[1] Then a post-Hoc analysis was performed on the results from the first half of the test to see which means could be compared. Its results are shown in Fig. B.6. In this chart, environment 1 is the De Jong recording, environment 2 is the recording convolved with omnidirectional directivity, and environment 3 is the environment convolved with BYU's directivity. The item listed as "Quality Place" refers to a metric created from the data, wherein the highest rating of the three environments received a score of 1, the second highest a score of 2, and the lowest quality received a score of 3. This chart shows

[1] A software package called Statistical Package for Social Scientists (SPSS) was used.

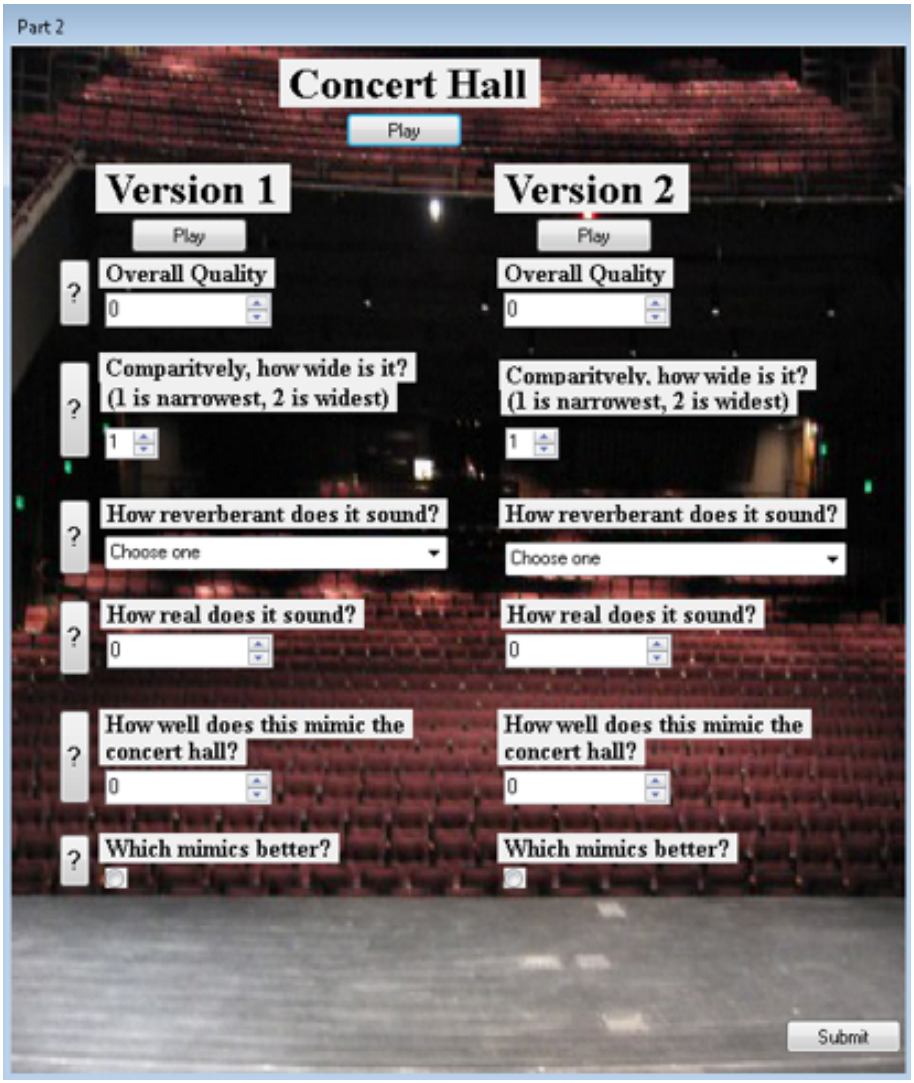


**Figure B.4** This is the second part of the test participants performed.

**Report**

| Environment | | Quality | QualityPlace | Width | Reverb | Real | Mimic |
|---|---|---|---|---|---|---|---|
| 1 | Mean | 6.79 | 1.37 | 1.39 | 2.72 | 7.47 | |
| | N | 76 | 76 | 76 | 76 | 76 | |
| | Std. Deviation | 1.835 | .670 | .655 | .624 | 1.807 | |
| 2 | Mean | 5.15 | 1.87 | 1.71 | 3.66 | 5.49 | 5.03 |
| | N | 152 | 152 | 152 | 152 | 152 | 72 |
| | Std. Deviation | 1.830 | .668 | .687 | .813 | 1.695 | 1.482 |
| 3 | Mean | 4.90 | 2.09 | 2.09 | 4.09 | 5.24 | 5.21 |
| | N | 152 | 152 | 152 | 152 | 152 | 72 |
| | Std. Deviation | 1.918 | .684 | .740 | .805 | 1.845 | 1.652 |
| Total | Mean | 5.38 | 1.86 | 1.80 | 3.64 | 5.79 | 5.12 |
| | N | 380 | 380 | 380 | 380 | 380 | 144 |
| | Std. Deviation | 1.994 | .723 | .749 | .920 | 1.968 | 1.567 |

**Figure B.5** The mean values gathered from the test.

that the means between either of the convolved recordings and the De Jong recording are statistially significant, but that the means of the quality values and real values are not statistically significant between the two convolved recordings. This signifies that only those two means cannot be compared.

Looking at the means, they indicate thatl the omnidirectional directivity convolved files generally had a higher quality place score than the BYU's direcitivty convovled files. However, the latter boasted more reverberance and greater ASW. The models were both reported to have less quality than the live recording, which is often true when trying to compare live concert halls to the models that represent them. The real recording also was reported to sound much more realistic than the files created by the model, which helps support that the listeners were correctly reporting realism.

An ANOVA test was also conducted on the second half of the test to see if the same patterns continued.[2] Its results are in Fig. B.7. This shows that for the second test, all the relationships between the means can be trusted except for the mimic values. The test was thus unable to predict which directivity model was able to mimic the live De Jong concert

[2] A Post-hoc analysis could not be used because the it requires at least three sets to compare.

**Multiple Comparisons**

| Dependent Variable | (I) Environment | (J) Environment | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower Bound | Upper Bound |
| Quality | 1 | 2 | 1.638* | .262 | .000 | 1.12 | 2.15 |
| | | 3 | 1.888* | .262 | .000 | 1.37 | 2.40 |
| | 2 | 1 | -1.638* | .262 | .000 | -2.15 | -1.12 |
| | | 3 | .250 | .214 | .244 | -.17 | .67 |
| | 3 | 1 | -1.888* | .262 | .000 | -2.40 | -1.37 |
| | | 2 | -.250 | .214 | .244 | -.67 | .17 |
| QualityPlace | 1 | 2 | -.500* | .095 | .000 | -.69 | -.31 |
| | | 3 | -.724* | .095 | .000 | -.91 | -.54 |
| | 2 | 1 | .500* | .095 | .000 | .31 | .69 |
| | | 3 | -.224* | .077 | .004 | -.38 | -.07 |
| | 3 | 1 | .724* | .095 | .000 | .54 | .91 |
| | | 2 | .224* | .077 | .004 | .07 | .38 |
| Width | 1 | 2 | -.316* | .099 | .001 | -.51 | -.12 |
| | | 3 | -.697* | .099 | .000 | -.89 | -.50 |
| | 2 | 1 | .316* | .099 | .001 | .12 | .51 |
| | | 3 | -.382* | .081 | .000 | -.54 | -.22 |
| | 3 | 1 | .697* | .099 | .000 | .50 | .89 |
| | | 2 | .382* | .081 | .000 | .22 | .54 |
| Reverb | 1 | 2 | -.941* | .109 | .000 | -1.16 | -.73 |
| | | 3 | -1.362* | .109 | .000 | -1.58 | -1.15 |
| | 2 | 1 | .941* | .109 | .000 | .73 | 1.16 |
| | | 3 | -.421* | .089 | .000 | -.60 | -.25 |
| | 3 | 1 | 1.362* | .109 | .000 | 1.15 | 1.58 |
| | | 2 | .421* | .089 | .000 | .25 | .60 |
| Real | 1 | 2 | 1.980* | .250 | .000 | 1.49 | 2.47 |
| | | 3 | 2.230* | .250 | .000 | 1.74 | 2.72 |
| | 2 | 1 | -1.980* | .250 | .000 | -2.47 | -1.49 |
| | | 3 | .250 | .204 | .221 | -.15 | .65 |
| | 3 | 1 | -2.230* | .250 | .000 | -2.72 | -1.74 |
| | | 2 | -.250 | .204 | .221 | -.65 | .15 |

*. The mean difference is significant at the 0.05 level.

**Figure B.6** This is the table generated by SPSS for the Post-hoc analysis of the data gathered from the first test. Environment 1 is the De Jong recording, environment 2 is the recording convolved with omnidirectional directivity, and environment 3 is the environment convolved with BYU's directivity

**ANOVA**

| | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| Quality | Between Groups | 193.761 | 2 | 96.880 | 27.803 | .000 |
| | Within Groups | 1313.671 | 377 | 3.485 | | |
| | Total | 1507.432 | 379 | | | |
| QualityPlace | Between Groups | 26.563 | 2 | 13.282 | 29.152 | .000 |
| | Within Groups | 171.763 | 377 | .456 | | |
| | Total | 198.326 | 379 | | | |
| Width | Between Groups | 26.668 | 2 | 13.334 | 27.008 | .000 |
| | Within Groups | 186.132 | 377 | .494 | | |
| | Total | 212.800 | 379 | | | |
| Reverb | Between Groups | 94.066 | 2 | 47.033 | 78.121 | .000 |
| | Within Groups | 226.974 | 377 | .602 | | |
| | Total | 321.039 | 379 | | | |
| Real | Between Groups | 274.224 | 2 | 137.112 | 43.331 | .000 |
| | Within Groups | 1192.934 | 377 | 3.164 | | |
| | Total | 1467.158 | 379 | | | |
| Mimic | Between Groups | 1.174 | 1 | 1.174 | .476 | .491 |
| | Within Groups | 349.819 | 142 | 2.464 | | |
| | Total | 350.993 | 143 | | | |

**Figure B.7** An ANOVA of the second half of the test

hall better.

Another analysis was an inspection of the impact of past mixing experience. An ANOVA table was created using the question regarding past mixing experience as the independent variable (Fig. B.8). This provided very interesting results. Those with mixing experience gave statistically different quality values, but (as shown by the *p*-value for quality place) acted very similarly to those without mixing experience. This is similar to past research done by Soren Bech [16] wherein he also found that mixing experienec and training impacts the quality ratings, but does not impact the preferences of acoustic sources. It also shows that those with mixing experiences perceived the realism of the different files very differently than those without mixing experience. Further analysis could have been done using only those with mixing experience to see if they were better able to distinguish which of the two directivity models produced a more realistic source.

| ANOVA – Is there past mixing experience? | | | | | | |
|---|---|---|---|---|---|---|
| | | Sum of Squares | df | Mean Square | F | Sig. |
| Quality | Between Groups | 15.074 | 1 | 15.074 | 3.818 | .051 |
| | Within Groups | 1492.358 | 378 | 3.948 | | |
| | Total | 1507.432 | 379 | | | |
| QualityPlace | Between Groups | .070 | 1 | .070 | .134 | .715 |
| | Within Groups | 198.256 | 378 | .524 | | |
| | Total | 198.326 | 379 | | | |
| Width | Between Groups | .000 | 1 | .000 | .000 | 1.000 |
| | Within Groups | 212.800 | 378 | .563 | | |
| | Total | 212.800 | 379 | | | |
| Reverb | Between Groups | .092 | 1 | .092 | .108 | .743 |
| | Within Groups | 320.948 | 378 | .849 | | |
| | Total | 321.039 | 379 | | | |
| Real | Between Groups | 47.235 | 1 | 47.235 | 12.575 | .000 |
| | Within Groups | 1419.923 | 378 | 3.756 | | |
| | Total | 1467.158 | 379 | | | |
| Mimic | Between Groups | .334 | 1 | .334 | .135 | .714 |
| | Within Groups | 350.659 | 142 | 2.469 | | |
| | Total | 350.993 | 143 | | | |

**Figure B.8** An ANOVA with past mixing experience as the independent variable

# Appendix C

# CVR One-Tailed Test

This appendix contains a table that outlines the critical values of essential votes when computing CVR values. It relates the panel size to the proportion of the groups that must rate the CVR as being essential, what the exact critical CVR score is, its one-sided $p$-value, and the critical number of essential votes. It is pulled from work by Ayre and Scally [6].

**Table 1.** $CVR_{critical}$ One-Tailed Test ($\alpha = .05$) Based on Exact Binomial Probabilities.

| N (Panel Size) | Proportion Agreeing Essential | $CVR_{Critical}$ Exact Values | One-Sided p Value | $N_{critical}$ (Minimum Number of Experts Required to Agree Item Essential)—Ayre and Scally, This Article | $N_{critical}$ Calculated From CRITBINOM Function—Wilson et al. (2012) |
|---|---|---|---|---|---|
| 5 | 1 | 1.00 | .031 | 5 | 4 |
| 6 | 1 | 1.00 | .016 | 6 | 5 |
| 7 | 1 | 1.00 | .008 | 7 | 6 |
| 8 | .875 | .750 | .035 | 7 | 6 |
| 9 | .889 | .778 | .020 | 8 | 7 |
| 10 | .900 | .800 | .011 | 9 | 8 |
| 11 | .818 | .636 | .033 | 9 | 8 |
| 12 | .833 | .667 | .019 | 10 | 9 |
| 13 | .769 | .538 | .046 | 10 | 9 |
| 14 | .786 | .571 | .029 | 11 | 10 |
| 15 | .800 | .600 | .018 | 12 | 11 |
| 16 | .750 | .500 | .038 | 12 | 11 |
| 17 | .765 | .529 | .025 | 13 | 12 |
| 18 | .722 | .444 | .048 | 13 | 12 |
| 19 | .737 | .474 | .032 | 14 | 13 |
| 20 | .750 | .500 | .021 | 15 | 14 |
| 21 | .714 | .429 | .039 | 15 | 14 |
| 22 | .727 | .455 | .026 | 16 | 15 |
| 23 | .696 | .391 | .047 | 16 | 15 |
| 24 | .708 | .417 | .032 | 17 | 16 |
| 25 | .720 | .440 | .022 | 18 | 17 |
| 26 | .692 | .385 | .038 | 18 | 17 |
| 27 | .704 | .407 | .026 | 19 | 18 |
| 28 | .679 | .357 | .044 | 19 | 18 |
| 29 | .690 | .379 | .031 | 20 | 19 |
| 30 | .667 | .333 | .049 | 20 | 19 |
| 31 | .677 | .355 | .035 | 21 | 20 |
| 32 | .688 | .375 | .025 | 22 | 21 |
| 33 | .667 | .333 | .040 | 22 | 21 |
| 34 | .676 | .353 | .029 | 23 | 22 |
| 35 | .657 | .314 | .045 | 23 | 22 |
| 36 | .667 | .333 | .033 | 24 | 23 |
| 37 | .649 | .297 | .049 | 24 | 23 |
| 38 | .658 | .316 | .036 | 25 | 24 |
| 39 | .667 | .333 | .027 | 26 | 25 |
| 40 | .650 | .300 | .040 | 26 | 25 |

**Figure C.1** A table that shows the critical CVR values [6].

# Appendix D

# Standards

Figure D.1 contains a list of different standards that exist for measuring subjective quality in different sources. Figure D.2 comes from Ref. [20], pages 242-243. This table gives the basic requirements for room considerations of five major standards and can serve as a reference point for room design and set-up.

The International Telegraph Union Telecommunication Sector (ITU-T) and the International Telegraph Union Radio Communication Sector (ITU-R) have extensive lists of different standards. Their names are listed in the following sections. The information was obtained through Ref [20], pages 10 and 11.

**ITU-T Standards**

**General Guidance**

- Handbook of telephonometry
- P.800 Methods for subjective determinatino of transmission quality
- P.800.1 Mean Opinion Score (MOS) terminology

| Name of Research Group | Standard Code | Standard Name |
|---|---|---|
| Audio Engineering Society | AES20 | AES recommended practice for professional audio - Subjective evaluation of loudspeakers |
| American National Standards Institute | ANSI S3.1-1999 (R2003) | Maximum permissible ambient noise levels for audiometric test rooms. |
| | ANSI S3.2-1989 (R1999) | Method for measuring the intelligibility of speech over communications system. |
| European Broadcasting Union | Tech 3276 | Listening conditions for the assessment of sound program material: monophonic and two-channel stereophonic |
| | Tech 3276: Supplement 1 | Listening conditions for the assessment of sound programme material: multichannel sound |
| | Tech 3286 | Assessment methods for the subjective evaluation of the quality of sound programme material - Music |
| | Tech 3286: Supplement 1 | Assessment methods for the subjective evaluation of the quality of sound programme material - Multichannel |
| International Electrotechincal Commission | 60268-13 | Sound system equipment - Part 13: Listening tests on loudspeakers |
| The international Telecommunications Union Standards | ITU-T | |
| | ITU-R | |

**Figure D.1** A list of useful standards for subjective evaluations [20].

| | IEC 60268-13 | ITU-R BS 1116-1 | EBU 3276 | AES 20 | N 12-A |
|---|---|---|---|---|---|
| Application | Listening tests of loudspeakers in domestic environments | Subjective assessment of small impairment | Critical assessment and selection of programme material | Listening tests of studio and high-quality loudspeakers | Reference listening room for listening tests |
| Basis | - | Expansion of IEC 60268-13 | Similar to ITU-R BS.1116-1 | - | - |
| Floor area (m^2) | 1-2 channel: 25-40<br>Multichannel: 30-45 | 1-2 channel: 25-60<br>Multichannel: 30-70 | >40 | >20 | 60 +- 10 |
| Room volume (m^3) | - | - | <300 | 50-120 | - |
| Aspect ratio | (w/l) | | | | |
| Reverberation time (s) | | | | | |
| Early energy | - | 10 dB attenuation of early reflections (15 ms, 1 - 8 kHz) | 10 dB attenuation of early reflections (15 ms, 1 - 8 kHz) | - | 15 dB attenuation of early reflections (10 ms, >400 Hz) |
| Late energy | - | - | - | Suppress flutter echoes | Sufficient diffusion over listening area to avoid flutter echoes |
| Background noise level | NR15 | NR10, NR15 max | NR10, NR15 max | 35 dBA and 50 dBC | NR10 or 15 dBA |
| Loudspeaker issues | 1-2 channel and multichannel | 1-2 channel and multichannel | 1-2 channel and multichannel | - | 1-2 channel only, refer to Recommendation N12-B for loudspeaker requirements |
| Headphone issues | - | Diffuse-field frequency response according to ITU-R BS.708 | Meet frequency response requirements of ITU-R BS.708 and otherwise meet requirements of IEC 60581-10. (Applied to mono and stereo only) | - | - |
| Listener issues | - | - | - | - | Capacity: 6-10 listeners |

**Figure D.2** A list of requirements for room characteristics in different standards. Based on work by Bech [20].

**Listening Test Methods**

- P.84 Subjective listening test method for evaluating digital circuit multiplication and packetised voice systems
- P.85 A method for subjective performance assessment of the quality of speec voice output devices
- P.830 Subjective performance of telephone-band and wideband digital codes
- P.831 Subjective performance evaluation of network echo cancellers
- P.832 Subjective performance evaluation of hands-free terminals
- P.835 Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm
- P.840 Subjective listening test method for evaluating circuit multiplicaiton equipment
- P.851 Subjective quality evaluation of telephone services based on spoken diallogue systems
- P.880 Continuous evaluation of time-varying speech

**Objective Models**

- P.563 Single-ended method for objective speech quality assessment in narrowband telephony applications
- P.862 Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs

- P.862.1 Mapping function for transforming P.862 raw results to MOS-DQO
- P.862.2 Wideband extension to Recommendation P.862 for the assessment of wideband telephone networds and speech codecs
- P.862.3 Application guide for objective quality measurement based on Recommendations P.862, P.862.1, and P.862.2

**Audio-visual**

- P.910 Subjective video quality assessment methdos for multimedia applications
- P.911 Subjective audiovisual quality assessment methods for multimedia applications
- P.920 Interactive test methods for audiovisual communications

**ITU-R Standards**

**General Guidance**

- BS.1283 A guide to ITU-R recommendations for subjective assessment of sound quality
- BS.1284 General methods for subjective assessment of sound quality

**Listening and Methods**

- BS.1116-1 Methods for subjective assessment of small impairments in audio systems including multihannel sound systems
- BS.1265 Pre-selection methods for subjective assessment of small impairments in audio systems

- BS.1534 Method for subjective assessment of intermediate quality levels of coding systems
- BS.1679 Subjective assessment of the quality of audio in large screen digital imagery applications intended for presentation in a theatrical environment

**Objective Models**

- BS.1387-1 Method for objective measurement of audio quality

**Audio Visual**

- BS.500-11 Methodology for subjective assessment of the quality of television pictures
- BS.7751 Multichannel stereophonic sound system with and without accompanying picture
- BS.1286 Method for subjective assessment of audio systems with accompanying pictures

# Bibliography


[1] J. Atkinson, "Interview of J. Gordon Holt," Stereophile (November, 2007).

[2] F. Toole, *Sound reproduction: The acoustics and psychoacoustics of loudspeakers and rooms* (Elsevier, Amsterdam, 2008).

[3] B. Report, "The Belmont Report: Ethical principles and guidelines for the protections of human subjects of research," http://www.intel.com/technology/silicon/lithography.htm (Accessed January 15, 2015).

[4] "Image pulled from Creative Commons," .

[5] B. Brown, S. Hendrix, D. Hedges, and T. Smith, *Multivariate analysis for the biobehavioral and social sciences* (Wiley and Sons, New York, 2012).

[6] C. Ayre and A. Scally, "Critical values for Lawshe's content validity ratio: Revisiting the origional methods of calculation," Sage 47, 79-86 (2013).

[7] A. E. S. Standards, "AES20," .

[8] S. Bech, "Perception of timbre or reproduced sound in small rooms: Influence of room and loudspeaker position," Journal of the Audio Engineering Society 42, 999-1007 (1994).

[9] S. Olive and P. Schuck, "The variability of loudspeaker sound quality among four domestic-sized rooms," Presented at the 99th convention of the Audio Engineering Society Preprint (1995).

[10] F. Toole, "Loudspeaker measurements and their relationship to listener preferences," Journal of the Audio Engineering Society 34, 227-235 and 323-348 (1986).

[11] Y. Ando, "Subjective preference in relation to objective parameters of music sound fields with a single echo," Journal of the Acoustical Society of America 24, 14-19 (1977).

[12] S. Olive, "A method for training listeners and selecting program material for listening tests," presented at the 97th Convention of the Audio Engineering Society (1994).

[13] P. D'Antonio, *Acoustic absorbers and diffusers* (Taylor and Francis, London, 2008).

[14] F. Toole and S. Olive, "Hearing is believing vs. believing is hearing: Blind vs. sighted listening tests and other interesting things," Audio Engineering Society Preprint 122-142 .

[15] A. Gabrielsson, B. Hagerman, and T. Bech-Kristensen, *Assessment of perceived sound quality in high fidelity systems* (Karolinska Institute, Stockholm, 1979).

[16] S. Bech, "Selection and training of subjects for listening tests on sound-reproducing equipment," Journal of the Audio Engineering Society 40, 590-610 (1992).

[17] S. Olive, "Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: A case study," Journal of the Audio Engineering Society 51, 806-825 (2003).

[18] A. Behrman, *Speech and voice science* (Plural Publishing, San Diego, CA, 2013).

[19] M. Nilsson, S. Soli, and J. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and noise," Journal of the Acoustical Society of America 92(2), 1085-1099 (1994).

[20] S. Bech and N. Zacharov, *Perceptual audio evaluation: Theory, method and application* (John Wiley and Sons, Chichester, England, 2008).