8.7 Auditory experiments

The predominant part of this book discusses the function of the electric guitar by way of physical laws, documented via formulae and measurement protocols. This enables us to explain e.g. wave propagation, induction and signal filtering – but not the actual effect on the listener. The verdict of the latter is only made available in auditory experiments. Therefore, the following seeks to give a short summary of methods towards controlled sound appraisal.

8.7.1 Psychometrics

Psychophysics forms an interdisciplinary scientific area bridging **psychology** (= the science of sensory perception, among others), and **physics** (= the science of natural processes); it researches and describes the connection between physical stimuli on the one hand, and the sensations and perceptions caused by these stimuli on the other hand. **Psychoacoustics** narrows the wide area of physics down to sound phenomena, and connects the "science of sound" with the "science of hearing". **Psychometrics** is a sub-area of psychology that has specialized in the (in particular quantitative) measurement of sensations. Electrical voltage is measured with a voltmeter, temperature is measured with a thermometer – but how can we measure the sensation of sound resulting from listening to a guitar? This can work only if the human being is both **measurement object** and **measurement device**, with all connected problems. The human being is the measurement object because his/her sound-perceptions are to be determined; and he/she is the measurement device because he/she needs to describe these perceptions. Since measurement object and measurement device cannot be separated, errors are possible. The statement "I do not hear any tone" can mean that the measurement object (the "subject") indeed does not hear anything and responds truthfully. However, it could also mean that the subject lies and does actually hear something. It could also indicate, though, that the subject thinks that what he/she hears is not a tone but e.g. a noise – in this case the response "not … any tone" would be truthful from his/her perspective. In order to avoid such misunderstandings, and to obtain the subject's assessment in the most unaffected and most reproducible manner, psychometrics has elaborated guidelines for the execution of experiments and their evaluation.

Reproducibility of the sound-presentation constitutes a particularly essential aspect. The reason that a guitar sounds – compared to the studio – different on stage is found more in the (physical) room acoustics, and not primarily in perception psychology, although the assessment *criteria* (measurement device!) can be situation-dependent, as well. In order to guarantee reproducibility in the presentation, many experimenters used specially equalized **headphones**. While this is an improvement over exposing the test person to a totally undefined sound field, it does not warrant an exact sound exposure, either. The position of the headphone (relative to the external ear), and the individual shape of the earlobe and the ear canal do influence the sound level.[♣] Another problem is the fact that an entirely unnatural sound field is created that turns with the head. Using precise instructions, mechanical fixation, probe microphones, and figurative presentations, these uncertainties can be reduced to the point that they are seen as "bearable" in daily research routine – this is then simply is as good as it gets. Sound presentation via one or two **loudspeakers** would be the alternative – not small PC-monitors, though, but calibrated premium studio monitors. Indispensable is again documentation: room acoustics, transfer functions, impulse responses, best supplemented by dummy-head recordings. The more is documented, the easier the decision after an experiment series whether an effect is due to the hearing system or due to experimental methodology.

[♣] Zollner M.: Interindividuelle und intraindividuelle Unterschiede bei Kopfhörerdarbietungen, Cortex 1994.

It may be that not a stored (or artificially generated) sound is to be assessed, but a **sound source**, i.e. an acoustic guitar or a guitar loudspeaker. In this case the question should be considered whether a recording via microphone or dummy head is made (and the recording then is listened to as mentioned above) or whether live-presentation is preferred (incl. documentation like the one involving loudspeaker presentation). A curtain hung in front of a picture will change the visual perception, and similarly the room between sound source and listener will influence the auditory perception. If the filtering by the room is ignored, the assessment is unusable. Only after the sound presentation is fully optimized and documented, the assessment of the sound may be started.

Auditory experiments may be of simple or complex, of fundamental or special character. **Threshold measurements** are easy to do for the subjects. Psychometrics distinguishes between (absolute) thresholds of stimulation, and just noticeable differences. The **threshold of stimulation** tackles merely the issue of whether something is heard. Not every tone with a sound power different from zero is audible – to be heard, the tone-level needs to be above the threshold-level. The threshold of stimulation that is determined for tones that are presented *in quiet* is called the **threshold in quiet**. If another (interfering) sound is present besides the sound to be assessed, the term used is **masked threshold**, e.g. "threshold masked by pink noise". When determining **just noticeable differences**, the question is from which degree of signal change a subjective difference is noticeable. For example: which change in frequency is necessary so that a change in pitch is perceived? The subject's task becomes more difficult if the question is not just whether a change is heard but also how big this change is. This **magnitude estimation** targeting the numerical assessment of perceived difference can lead to significant scatter up to the point that it is actually impossible for some experiments. We can "force" assessments, but it is hardly measurable whether something sounds better by a factor of two or three. Psychoacoustics states that it is measurable whether a sound has double the loudness of another sound. Yeah, kind of – but with a scatter of ±6 dB, gripe the critics. Scatter of measurement results is not at all limited to psychometric experiments – all measurements will include variance. It's just that in psychometrics, the variances are particularly pronounced and therefore need to be looked into with particular scrutiny.

No subject will increase the level always by exactly 10 dB when asked to adjust to double the loudness. That is why the experimenter will average the intra-individually varying values, deriving a subject-specific mean value. *One* subject would represent an unsuitably small sample, and thus e.g. 24 further subjects need to do this adjustment-experiment, leading to 25 different **mean values** that show inter-individual differences. Again, an average is taken, and finally we get the result that will e.g. express that "on average" the subjects will increase the level by 10 dB to achieve double the loudness. That this mean is not valid for each and every human being – that is often pushed to the back of our minds. So let's play devil's advocate: literature reports scatter between 5 – 17 dB, and even 4 – 30 dB is found [Hellbrück 1993]. Even so: *here the center of the distribution was in the class of 8,6 – 9,8 dB*. Well then … that is almost 10 dB. To conclude from the variance that the whole shebang is one giant hokum – that would show some uncalled-for ignorance, after all. Insofar as experimenter and subject are aware of what they evaluate, averaging methods offer the only possibility to reduce clusters of dots to functions. Whether the assessments of fluctuation strength include a scatter of factor 4 or 8 – they still clearly feature a band-pass characteristic with a maximum at a modulation frequency of 4 Hz. We simply have to avoid the mistake to declare such results – with a three-digit precision – as universally valid; average values do have a limited accuracy, too.

Of course, experiment and averaging become questionable if experimenter and subject have different attributes in mind. A strongly exaggerated example would be the following: the experimenter distributes written instructions regarding the scaling of the sonority of drums. Questions are not allowed so as not to influence the subject. And off we go – judging away on a scale from 0 to 10. Not wanting – as a spoor student – to forgo those hourly €15.-, one tags along. Either according to the best of one's knowledge (or rather: perception), or according to the Monte-Carlo-method: everything's coming 'round again, and even this hour will pass. The PC generates some averages, and we have a result. The concept what "sonority" is supposed to be – that should be shared by experimenter and subject … otherwise it all really is one big hokum. And nobody say that a good result proves that this term "sonority" is self-explanatory.

A less construed example from the Süddeutsche Zeitung (an internationally read German newspaper) published on 24.09.2009: positioned within an MRI scanner, a subject is shown various photographs. Depending on the motif, the MRI scanner establishes different brain activities. Exceptional here: the subject is a fish. And even much more exceptional: the fish is dead. In spite of this, the evaluating computer manages to arrive at a significant mean result. In this case, the experimenter is not a charlatan but an honorable scientist seeking to show *how much nonsense is often practiced in modern experimental brain research*. N.B.: having many subjects at hand and using modern ("Russian") averaging algorithms won't guarantee solid data … or, in other words: garbage in – garbage out.

Modern psychology, and in particular psychometrics, increasingly employs statistical evaluation methods; that may be pesky, but it's unavoidable. The most wonderful experiment is no good if the results are erroneously evaluated. Just as nonsensical is to continue to (without experimental experience) process mindless data until a convenient result is obtained. Consider that, in a source-recognition experiment, all guitars are given the numeral 1, all trombones the number 2, and all basses the numeral 3. If the subject has now recognized four times the 1, twice the 2, and four times the 3, then we may not average arithmetically and state that as a mean value a trombone is recognized. These assessments or nominal judgments, after all, and there is no mean value. It would be similarly absurd to calculate a "mean postal area code". That would be possible, yes, but not interpretable.

A **nominal judgment** groups according to names and thus congregates elements of equal attributes into groups. Only with an **ordinal judgment**, a ranking is created – however without any metric. In metrology, class-0 is more precise than class-1, and the latter is again more precise than class-2. Class-0, however, does not necessarily feature double the precision of class-1, and if that were the case, class-1 could well be 3 times as precise as class-2. More mathematically: an ordinal scale is determined via inequations but not via intervals of equal size. The latter comes into play only with **interval scales**, they allow for additivity based on equidistance. What is not required is that the property of the element with the value "0" disappears. 0°C does not imply "no temperature" but rather is an arbitrarily fixed neutral point, and that is also why 20°C is not double as warm as 10°C. At the end of this list we have the relational scale in which the relations of the numbers mirror the relation of the degree of manifestation of the assessed characteristics. The sone-scale is such a relational scale: if two loudnesses have the relation 2:1, the same ratio is also found in the corresponding sone-numbers (8 sone is double the loudness of 4 sone). Conversely, the phon-scale is not a relational scale: 60 phon is not double the loudness of 30 phon.

The following table summarizes scales, properties and operations. Nominal scaling only offers *equal* or *unequal*, ordinal scaling adds in *larger than* and *smaller than*, additivity comes in with the interval scale, and product/division is only there from the relational scale.

The median (numerical value) of a nominally scaled set cannot be determined because for this all elements need to be brought into a ranking – which does not exist in nominal scaling. Only the modus, the maximum rate of occurrence, may be identified. “Most letters were transported for postal code 93057” makes sense, but “the median is postal code 93057” does not. As a rule, to use ratios of levels is pointless – although there may be exceptions here and there, insofar as “0 dB” indeed is mean to imply “nothing”. In terms of the SPL, level ratios are usually without meaning – using an equalizer, however, a boost of 8 dB may be double the boost of 4 dB.

Scale	**Nominal**	**Ordinal**	**Interval**	**Relational**
Synonyms	Topologic scale		Metric scale, cardinal scale	
Allowable statistical measures	Absolute and relative rate of occurrence, modus	Cumulative rate of occurrence, median, percentile	Arithm. mean value, variance, standard deviation	Geometric mean value
Operations	$= \neq$	$= \neq, <, >$	$= \neq, <, >, +, -$	$= \neq, <, >, +, -, \times, \div$
Features	Nominal feature, categorical or qualitative feature	Ordinal feature, ranking feature, comparative feature	Cardinal feature, quantitative or metric feature	

Table: Scales, features, allowable operations. In addition to the statistical measures in each column, all measures on the left of these are, correspondingly, also allowed.

Once we now have perfected the sound to be presented, and once the feature-scale to be found is determined, the subjects (test persons) may arrive. From now on it’s: no influencing, and reproducible instructions. With a statement given right at the start of the sort that EC’s “Brownie” is to be assessed, an opinion like “sounds a bit thin” is not likely to be voiced – that guitar will simply sound “killer”. In order to prevent such bias, the desired objective is the **blind test**, although that is not always doable. It would be possible to assess two guitar amps without prejudice if the amps are hidden behind an opaque curtain (a rotary table takes care of positioning problems); however, the immediate difference between a Gibson Les Paul and an ES-335 may only be hidden from the guitarist if rather elaborate precautions are taken. The differences between different scale lengths (e.g. 24" vs. 25,5") are always recognized – blind tests are impossible here. **Written** instructions for all subjects ensure that everyone is told the same, and they also facilitate checking the instructions a year later. If we realize in the course of an investigation that the subject have difficulties doing an assessment, we must not change the instructions until the ”correct” result turns up and average subsequently over all experiments. Out of the question is also something like averaging only over the last five subjects (because only they have heard the difference). Difficult question: should one single out unsuitable subjects? To assess drumsticks, you would not ask harpists to give a verdict; the sound of a guitar amplifier can, however, certainly be judged by a non-musician, as well. Because there are no set rules here, documentation is particularly essential (questionnaires handed out to all subjects). If we want to do a true service to science, we measure the hearing threshold in quiet (audiogram) of the subjects ahead of the start of the experiments. This is because many a musician (and other people spending any length of time in noisy environments) have generated (and have been subject to) so much sound energy in the course of their lives that their auditory system has experienced considerable damage. Corresponding judgments may therefore not be typical for those of normal hearing. Wouldn’t you concur with that, dear Mr. Townshend? Mr. Townshend, sir? Mr. Peter Townshend – HELLO there?? **MR. TOWNSHEND!!!**

Last, we have to consider according to which method the subject is going to deliver the judgment. That is, "last" in the framework of this short overview, because the rules of professional psychometrics[♣] are more extensive and go beyond the presently set scope. **Methods of acquiring judgments** differ (among other aspects) according to the degree of involvement of the subject. Is the latter merely supposed to give a verbal assessment ("I don't hear anything"), or does he/she need to twist a knob such that a tone becomes just audible (or inaudible)? Is a scale of the assessment presented, or can the subject make one him/herself? Is the verdict "no difference" allowed, or is a preference forced (forced choice)? Is the response of the subject considered when new test sounds are selected? May the subject compare test sounds as long as he/she likes, or is a decision called for after two repetitions? For decades, psychologists have never grown tired of preaching that all these details in the experiments are vital to the results, and so we engineers cannot but believe it, and promote it. All the while hoping that – vice versa – the advantages of correct level-measurements find a similarly strong lobby in the psychologist-camp.

Scientific auditory experiments are more than just calling in three pals to in order to verify the hypothesis that the new Fender is another milestone in rock history. The last trap is found in the formulation of the results. The statement "the Makkashitta VR-6 has some mighty sustain" is o.k.; however, declaring "due to its maple neck, the Makkashitta VR-6 has some mighty sustain" is, most probably, rubbish. Unfortunately, it is everyday practice in test reports: the tester hears something (which is his god-given right), and connects without any prove what he has heard to some kind of material characteristic (which is stultification of the reader). Often, e*vid*ent associations (i.e. from visual domain) are dragged into the arena in order to substantiate "ear-sounding" connections (i.e. in the auditory domain). Does a silver trumpet ring more "silvery" than a "warm-sounding golden trumpet? Science says: no, it's all but imagination, or influencing the player. If the latter has to play under yellow lights and cannot distinguish the metals, he/she plays the same, and then the sound is the same, too – despite different metals (and given equal geometry). Does that big loudspeaker have less treble because its heavy membrane is set in motion more slowly? Mechanics say: no, you are mistaking cutoff-frequency with efficiency. Are the sound pressures arriving at the two ear canals indeed the only excitation quantities for the auditory sense? Well, with the answer "of course not", the examinee would have most likely failed the psychoacoustics exam in 1979. But since then, much has progressed; we do learn all the time. The visual impressions play an important role in the auditory perceptions, and thus the perceived loudness is dependent on the distance at which we see the sound source. It's also why the red express train is perceived to be louder than the green one, despite equal SPL [Fastl]; and it is the reason why we may hear "behind us" although the sound source is in front. It's a wide field, and – for the most part – still an only sketchily examined one.

[♣] e.g.: Kompendium Hörversuche in Wissenschaft und industrieller Praxis, www.dega-akustik.de

