LUND UNIVERSITY

ENGINEERING ACOUSTICS, LTH
TVBA-1012, SWEDEN 2005

DRUM SOUND FROM FLOOR COVERINGS

– Objective and Subjective Assessment

ANN-CHARLOTTE JOHANSSON

ENGINEERING ACOUSTICS, LTH
CODEN: LUTVDG/(TVBA-1012) / 1-130 / (2005)
ISBN: 91-628-6531-5 ISSN: 0281-8477

DRUM SOUND FROM FLOOR COVERINGS

– Objective and Subjective Assessment

ANN-CHARLOTTE JOHANSSON

MED STÖD AV
STIFTELSEN FÖR KUNSKAPS-
OCH KOMPETENSUTVECKLING

Abstract

Drum sound is the sound produced when an object, such as a foot, hits the flooring in the room in which the receiving ear is located. Drum sound has attracted interest in recent years, particularly due to an increased use of thin floating floor constructions, such as veneer or laminate flooring, which can produce loud and sharp walking sound. A prediction model of the subjective response, in a paired comparison test, to drum sound based on differences in objective measurements is developed. The difference in 10-percentile loudness, N_{10}, between two stimuli is shown to predict the subjective perceived disturbance better than, for example, A-weighted sound pressure level. A difference of about 8% in N_{10} resulted in 50% of the assessors noticing a difference.

A comparison of different existing approaches to analysing the result from a paired comparison test is made. The main focus is set on the basic models by Thurstone-Mosteller and Bradley-Terry and extensions of these concerning ties. Procedures for testing if the responses and calculated ranking values are statistically different are presented. The advantages and disadvantages of these methods are discussed. These methods are illustrated with examples from tests on drum sound from floor coverings.

A branch norm has been established for measuring drum sound on laminate floor coverings. The norm evaluates the subjective perception of the drum sound's loudness using the ISO tapping machine. A round-robin study of the norm is reported along with the results of a paired comparison listening test using the same floor coverings. General aspects of evaluation measures, tapping machines, test environments, etc., that need to be considered when measuring drum sound on various floor coverings are discussed. It is concluded that loudness as measured according to ISO 532B correlates the best with the subjective perception of the drum sound's loudness. The tapping machine can be used to excite hard floor coverings to produce the drum sound, but should be used with caution in studying low-level drum sounds due to the tapping machine's inherent mechanical noise.

Key words

Building acoustics, psychoacoustics, drum sound, walking sound, paired comparisons, listening test, sound quality, loudness, foot impact, binaural recording, laminate flooring, parquet flooring, veneer flooring, tapping machine, acoustic standard.

Acknowledgements

This thesis is a product of 'The Building and Its Indoor Environment' research school at Lund University. The research school is financed through the KK-Foundation (the Swedish Foundation for Knowledge and Competence Development). The financial support from Pergo Europe AB is also gratefully acknowledged.

I want to acknowledge some of the people who have helped me during the project: my supervisors Per Hammer and Erling Nilsson for their advice and encouragement, Jonas Brunskog for his valuable annoying questions and interest, Robert Månsson for his help with all the measurements and Bo Zadig for helping me print this work. Many thanks to Dag Glebe, Klas Hagberg, Per Hiselius, Sven Lindblad, Lars-Göran Sjökvist for their support and interesting discussions. Thank you all for keeping up the nice atmosphere at the division.

I also want to thank the people at Pergo Europe AB, especially Håkan Wernersson and Peter Ringö, for all their support. The members of the EPLF technical committee are also acknowledged for their cooperation in the development of the norm.

Finally, I wish to express my appreciation to all of my friends and family who have lent their ears to some grief here and there, especially Ulf for his love, patience and never-ending encouragement.

Dissertation

Summary of Included Papers

Paper A: Aspects on paired comparison models for listening tests

In many acoustic environments, for example, buildings or vehicles, as well as in product development, etc., there is a need to rank and classify sounds. A frequently used procedure is the paired comparison test. A number of ways to perform and analyse this test exist. In this paper a comparison of different existing approaches is made. The main focus is set on the basic models by Thurstone-Mosteller and Bradley-Terry. Extensions to both of the models, concerning ties, are presented along with a discussion of when they should be used. Thereafter, procedures to test whether the calculated ranking values are statistically different are presented. The advantages and disadvantages of these methods are discussed, and some examples are given which consider the responses from tests on drum sound from floor coverings. It is seen that the choice between these models is not crucial. Ties are generally recommended as they add information and can decrease the results' confidence intervals/regions. The model by Bradley-Terry and its extensions are recommended. However, if only scale values are requested, the treatments are somewhat similar in character and no ties are allowed, the Thurstone-Mosteller model is recommended due to the simplicity of the calculations.

Paper B: Prediction of subjective response from objective measurements applied to walking sound

The paper discusses prediction of the subjective response to walking sound — also called drum sound – based on differences in objective measurements. "Walking sound" refers to the sound heard when someone is walking in the same room as the listener. Walking sound has attracted interest in recent years, particularly due to an increased use of thin floating floor constructions, such as veneer or laminate flooring, which can produce loud and sharp walking sound. A paired comparison test was performed in laboratory where listeners were asked which of the walking sounds was most disturbing. The response was analysed using a modified Bradley and Terry model allowing ties. Various measures, such as loudness according to ISO 532B, were tested against the subjective response using linear regression. The difference in 10-percentile loudness, N_{10}, between two stimuli was shown to predict the subjective response better than, for example, A-weighted sound pressure level. A difference of about 8% in N_{10} resulted in 50% of the subjects noticing a difference. The methodology used is applicable in situations when objective measures that have subjective counterparts are sought. Although the method is based on relative observations, an absolute ranking can be

obtained by using a reference or a well-defined recording situation.

Paper C: Evaluation of Drum Sound with ISO Tapping Machine

A branch norm, EPLF NORM 021029–3, has been established for measuring drum sound on laminate floor coverings. "Drum sound" refers to the sound occurring when an object, e.g. a foot, strikes the flooring in the room in which the receiving ear is located. The norm evaluates the subjective perception of the drum sound's loudness using the ISO tapping machine. A round-robin study of the norm is reported along with the results of a paired comparison listening test using the same floor coverings. The article discusses general aspects of evaluation measures, tapping machines, test environments, etc., that need to be considered when measuring drum sound on various floor coverings, such as linoleum, wood parquet and laminate. It is concluded that loudness as measured according to ISO 532B correlates the best with the subjective perception of the drum sound's loudness. The tapping machine can be used to excite hard floor coverings to produce the drum sound, but should be used with caution in studying low-level drum sounds due to the tapping machine's inherent mechanical noise.

Introduction to the Field of Research

Contents

Chapter 1

Introduction

1.1 Background

Interest in the effect of background noise on health and work capacity has increased. One disturbing factor in office spaces, schools, hotels, etc., is the drum sound, i.e. the sound produced when an object, such as a foot, hits the flooring in the room in which the receiving ear is located, as in figure 1.1. Drum sound is sometimes also called "walking sound" or "drum noise", although drum noise should be avoided since the term "noise" suggests that the sound is unwanted, which is not necessarily the case.

The first study of drum sound known to the author was made in Denmark in 1952 by Larris [1]. In that study, measurements were made of the sound produced by the ISO tapping machine on various types of flooring, and listening tests were performed in which assessors judged the level of the sound of a person walking on the floor, with the help of a Barkhausen phon-meter (generating an 800 Hz tone). The first reference to drum sound in Sweden was made in 1958 by Brandt [2]. The interest in performing further investigations in the field of drum sound seemed, however, to be low for many years. Nevertheless, drum sound has attracted interest in recent years, particularly due to the increased use of thin floating floor constructions, such as veneer or laminate flooring, which can produce loud, sharp drum sounds when a person wearing hard-heeled shoes walks on them. Due to consumers' increasing demand for floorings with improved drum sound properties, industry has become interested in producing better products.

Improving a product's sound sometimes means lowering the sound pressure of the radiated sound. In general, however, improving the product's sound quality involves much more than simply lowering the sound pressure level. Product sound quality was defined by Jekosch and Blauert [3] as "a

Figure 1.1: Drum sound.

descriptor of the adequacy of the sound attached to a product." A motorbike should not sound like a vacuum cleaner; hence, it is not only by changing the perceived loudness that the product sound is improved. Other descriptors of the sound are needed. Psychoacoustics is defined as "the science which deals with the relationship between parameters of acoustic waves and attributes of auditory events" [4]. As psychoacoustic measures aim at describing the hearing sensation based on acoustic stimuli, these measures provide a useful tool when a product's sound quality is to be described and improved. By performing listening tests the hearing sensation of a sound is obtained. Various response scales are often used for evaluating sound quality. Another approach is the method of paired comparisons [5]. In such a test assessors are asked to tell which of two sounds has a certain attribute (such as a pleasant sound); in some tests the assessors are permitted to declare a tie (hence indicating that no difference is perceived). With the result from the listening tests, the available acoustic and psychoacoustic measures are then combined so that accurate descriptors of the hearing sensation of the sound are identified.

The drum sound study in [6, 7] is the starting point of this thesis. Recordings of a walking female and male on various floor coverings were used. In the study the assessors were asked in a paired comparison test which drum sound is the least disturbing and most pleasant sound. The assessors were also asked to assess the drum sounds by using scales of various adjectives such as perceived strength, pitch, hollowness, etc. (only presented in [6]). By calculating the correlation coefficient between the scaling of the adjec-

tives and the result of the paired comparison test, it was concluded that the perceived strength correlated well to the perceived pleasantness (correlation coefficient was -0.90 for the drum sounds created by the female and -0.92 for the "male drum sound"). The correlation coefficients were -0.87, -0.82 and -0.81 for hollowness, distinctness and sharpness respectively for the male drum sound whereas for the female they were lower (-0.63, 0.25 and -0.25). The results from the paired comparison tests were thereafter correlated to the objective (instrumental) results where, among others, the psychoacoustic measures loudness and sharpness were applied.

As floor coverings with better drum sound characteristics are being developed, the need for a standard method for measuring floor performance and presenting it to the market has arisen. A couple of measurement methods exist [8, 9, 10]. Most of them use the ISO standard tapping machine [11] as the sound source, but steel balls [12] and a real walker [13] are also used. There is a need to harmonise these methods. Naturally, the final method must be repeatable, reproducible and practical, and it must correspond to subjective perceptions of the sound.

1.2 Objectives

The first objective with this work is to find a prediction model of the subjective response to drum sound and to find out what difference in a sound is needed in order for the assessors to notice a difference. A methodology is to be presented that is applicable in similar situations when measures that have subjective counterparts are sought.

During the work, the large number of models for analysing the result from a paired comparison test were observed. As the listening tests are the key to finding how the sound should be improved it is important to know the advantages and disadvantages of various methods. The second objective is therefore to clarify the differences among the paired comparison models, especially with respect to whether and when ties should be allowed. Procedures for testing if the responses and calculated ranking values are statistically different are to be presented.

A branch norm, EPLF[1] NORM 021029–3 [14], has been established for measuring drum sound on laminate floor coverings. The division of Engineering Acoustics at Lund University has contributed to its development. The third objective is to discuss general aspects of evaluation measures, tapping

[1]EPLF is the association of European Producers of Laminate Flooring. Internet address: www.eplf.com

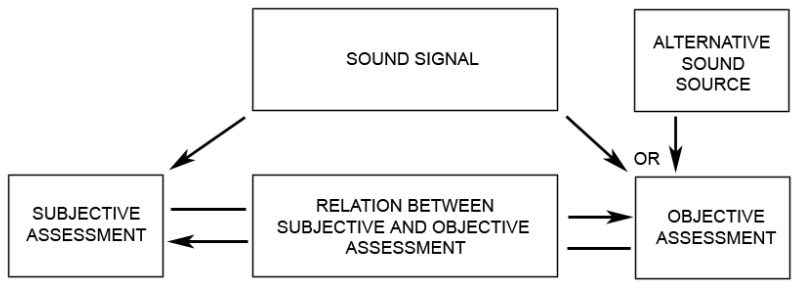

Figure 1.2: The methodology of the working process in papers B and C.

machines, test environments, etc., that need to be considered when measuring drum sound on various floor coverings, such as linoleum, wood parquet and laminate.

These three objectives correspond to the three papers [15, 5, 16] included in this thesis.

1.3 Method

In figure 1.2 the methodology of the work in papers B and C is illustrated. The methodology is, however, not restricted to drum sound but applies in situations where measures that have subjective counterparts are sought. The sound quality of the sound signal is evaluated in a subjective assessment where listening tests are performed. In the objective assessment instrumental measurements are performed either on the actual sound signal (as in paper B) or on an alternative sound source (as in paper C) to which various acoustic measures are applied. The relation between the subjective and objective assessments, which is the crucial part of all sound quality investigations, can thereafter be found.

The choice of method for the listening test is governed by the prerequisites and the objectives of the test. A brief description of the various alternatives is given in chapter 3. In this thesis, the method of paired comparisons is used. Aspects on this method are given in paper A [5]. In paper B [15] the drum sounds were assessed in the context of an office environment, where correlation between the surface and the sound is assumed to be less important than for domestic floors. It has been noticed by the author that assessors, when asked to choose a solution for their home, prefer different sounds de-

pending on the design of the surface. In an office environment, however, the focus is on decreasing the disturbance produced by walking sound. Thus, assessors were requested to imagine themselves in an office space and, in a paired comparison listening test, to say which of the sounds was the most disturbing. In paper C [16] general aspects considering a standardised measurement method to evaluate drum sound from floor coverings are addressed. The question concerned the perceived loudness as the optimal drum sound character might differ for various floor coverings and applications.

"Objective assessment" here means the instrumental assessment including application of acoustic and psychoacoustic measures. In chapter 2 the most common measures are explained. The sound used is either the same sound (original sound) that is used in the subjective assessment or an alternative sound source. In paper B [15], the former is used to enable the creation of a measure that, based on the actual sound, can predict the subjective response. The drum sound that is used depends on several parameters, the most important of which appear in figure 1.3. The parameters in italic are varied. In paper C [16] an artificial sound source is used as the objective in that case was to find a standard measurement method, for which a real foot is not appropriate. The artificial sound source is the ISO tapping machine that was developed to be used for evaluation of impact sound (the transmitted sound from the tapping machine to the room below) [11]. The result of a round-robin test of the objective evaluation of various laminate floor coverings is reported.

The drum sound produced by the tapping machine is obviously not the same as that produced by an actual foot; nevertheless by the use of various measures good correlation to the subjective response to the drum sound from an actual foot can be achieved. Even though the actual sound source is used in the objective measurements it is important to make sure the evaluation method has a subjective counterpart. In chapter 4 various methods to reveal and obtain relations between the subjective response and the objective measures are described. In this work, linear regression is used.

1.4 Limitations

The drum sounds are here evaluated for an office environment. In other environments other aspects, such as the character of the sound, might be important. As shown in figure 1.3, the influence of walking speed and background noise have not been addressed. The influence of different room acoustics has been addressed to some extent in the objective evaluation using the tapping machine but not in the subjective evaluation. Mainly thin floor coverings

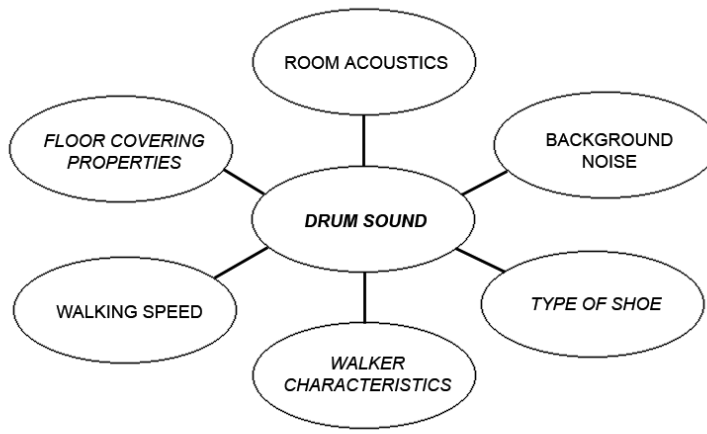

Figure 1.3: A complete model of how drum sound is perceived requires knowledge of several parameters. The italic parameters are investigated in the article.

on a thick homogeneous concrete subfloor have been evaluated, even though drum sound is also a problem for other constructions, such as installation floors.

Chapter 2

Objective assessment

To characterise the physical stimulus sound, the amplitude and pitch are usually given by the sound pressure level [dB] and frequency [Hz]. However, the physical stimulus is treated on its way through the hearing system so that the human perception of sound cannot simply be described in decibels and hertz. In the field of psychoacoustics the human perception of sound is investigated, and measures that better correspond to the hearing system are developed. Before some of these measures are discussed in detail, the drum sound generation and hearing system are briefly described.

2.1 Drum sound generation

Drum sound is produced when an object, such as a foot, hits the flooring in the room in which the receiving ear is located. The emitted sound consists partly of the sound created in the contact area where the foot (or hard heeled shoe) hits the floor and trapped air is pushed away, compare with a hand clap, and partly of the sound that the induced bending waves in the floor structure radiate to the surroundings. The radiated drum sound depends on many factors; the most important parameters appear in figure 1.3. Naturally the type of floor is one factor. In this thesis the construction of the floor is typically a thin, 7–14 mm, veneer or laminate floor covering with an underlayer, e.g. 2 mm polyethylene (PE) foam and a homogeneous (reinforced) concrete subfloor as shown in figure 2.1. Linoleum, textile and 22 mm wooden flooring on timber joists have been included as well, but only rarely.

The walker characteristics (weight, walking style) and type of shoe are another factors influencing the radiated sound. In figure 2.2 the time histories of the emitted drum sound pressure [Pa] of five steps are shown for three

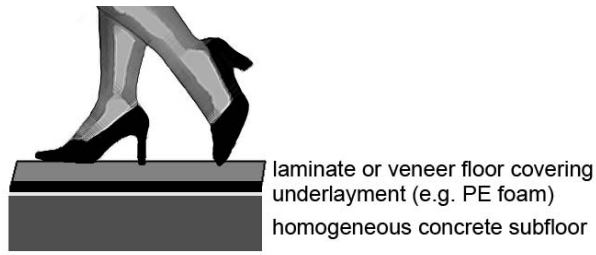

Figure 2.1: Typical construction of the floors.

cases: the first two are the results of the same female walking on two different floorings, i.e. a 22 mm wooden flooring on timber joists and a 14 mm veneer flooring on a soft PE foam. The third, lowest, is the result of a male walking on the 14 mm veneer flooring. The sound was measured when each person was walking towards and passing the microphone, which is why the amplitude is increasing for the first four steps and thereafter decreases. It can be seen that the female wearing high hard-heeled shoes is pushing the heel hard on the floor, whereas the male (wearing men's hard-heeled shoes) does not push the heel as hard. Another difference can be found comparing the amplitude of the first peak to the second peak where the front of the shoe hits the floor. A clear difference in the duration of the drum sound of a single impact can be seen between the 14 mm veneer flooring and the 22 mm wooden flooring on timber joists; resonances of the air cavity between the timber joists increase the duration of the sound.

The character of the sound that is radiated can be further examined in the frequency domain; see figure 2.3 where the equivalent sound pressure levels, SPL, for the original 3 s time history signals are displayed. F denotes female gait with high hard-heeled shoes, M denotes male gait with men's hard-heeled shoes. The floorings are 8 mm laminate flooring on fibreboard, 22 mm wooden flooring on timber joists and a 14 mm veneer flooring on a soft PE foam. It is seen that the character of the drum sounds differs mainly between about 80 and 8000 Hz. For 7–14 mm veneer or laminate floor covering on a concrete subfloor, the main differences between various floor coverings for the same walking person are usually between 200 and 8000 Hz.

Simple models of foot impact have been developed [17, 18]. The dynamic characteristics of a foot during impact is complicated. Measurements of the foot's mobility (velocity phasor response divided by the phasor of the exciting force) [m/s/N] wearing a shoe were made in [17, 19]. In figure 2.4 the results

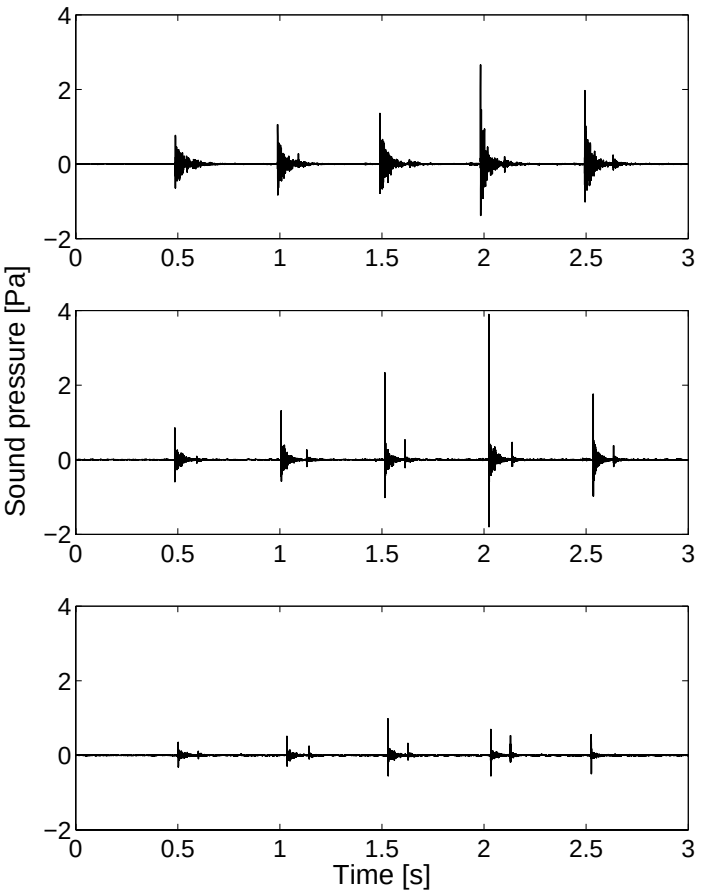

Figure 2.2: Time histories of emitted drum sound from five steps. The first two histories above are the result of the same female walking on two different floorings, i.e. a 22 mm wooden flooring on timber joists and a 14 mm veneer flooring on a soft PE foam. The third, lowest, is the result of a male walking on the 14 mm veneer flooring. The male and the female used different types of shoes.

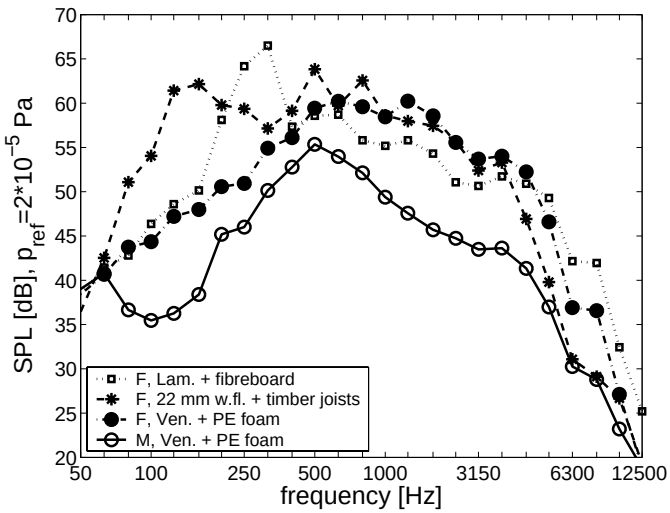

Figure 2.3: Various drum sounds in the frequency domain. F denotes female gait with high hard-heeled shoes, M denotes male gait with men's hard-heeled shoes. The floorings are 8 mm laminate flooring on fibreboard, 22 mm wooden flooring on timber joists and a 14 mm veneer flooring on a soft PE foam.

are given as the acceleration resulting from a constant alternating force, i.e. the input mobility times the radian frequency, $\omega = 2\pi f$ [rad/s], f is the frequency in Hertz; a constant-mass is then represented as a horizontal line. It is seen that the foot seems to behave like a mass of approximately 8 kg below about 10 Hz. Thereafter the behaviour is more like a spring, as the inclination is approximately ω^2. Above 200 Hz it behaves like a mass again but with a weight of approximately 40–150 g. The differences in the result might be due to different measurement set-ups and individual differences. In [20] an initial model of the interaction of the foot and the floor system is presented. The mobility of the foot is imbedded into the differential equation for the floor system. The interaction of the ISO standard tapping machine, which is used for measuring drum sound in [14, 16, 21], and the floor system is also presented. For the tapping machine one has to consider that the masses rebound, which can be simulated following the procedure in [22]. The study in [20] was an attempt to investigate the effect of cavities between the floor covering and the subfloor, however, more work is needed towards finding an adequate model of the interaction of the foot and floor system.

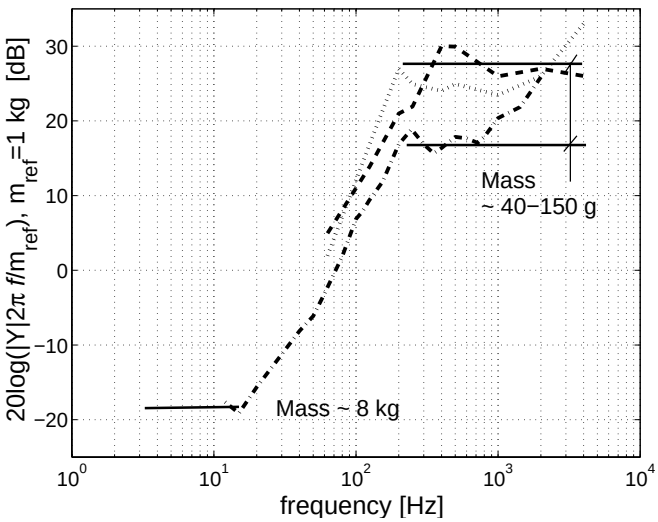

Figure 2.4: Measurements of mobility of a foot with and without shoe displayed as the relative acceleration with a reference of mass 1 kg. Dash-dotted line: female foot with high hard-heeled shoe mechanically "floated" by long nylon strings hung from the ceiling [17]; dashed line: on a shaker, standing male wearing a boot with rubber sole; and dotted line: sitting male without shoe [19].

2.2 Human auditory system

The human auditory system is complex. Comprehensive literature is available [23, 24] and the intention here is only to review some of the most important features of hearing.

The physiology of the human auditory system consists of three principal parts: the outer, middle and inner ear. The outer ear consists of the pinna and the ear canal. The pinna differentiates sounds from the front compared with those from the rear to some extent and it also works as an encoder of the direction of sound. The ear canal, about 2 cm in length, can be compared acoustically with an organ pipe, closed at the inner side by the eardrum, which causes a resonance effect at one-quarter of a wavelength. The sensitivity of the hearing system is therefore improved at about 4 kHz. In the middle ear the sound signal is amplified before it reaches the inner ear where the mechanical oscillations are transformed into nerve pulses. The transformation happens in the cochlea, which is filled with two different fluids and consists of three channels that run together and transmit the oscillations to the basilar membrane in between the canals. The basilar membrane supports the organ where the sensory hair cells are located. The length of the basilar membrane is about 32 mm. It is formed by thin elastic fibres tensed across the cochlea duct and the fibres are short and closely packed in the basal region and become longer and sparse proceeding towards the apex of the cochlea. The stiffness of the basilar membrane is therefore varied, which enables a sound signal to be picked up at different locations of the membrane, which, in turn, causes different nerve impulses to be transmitted to the brain.

A healthy young human ear can distinguish sounds with frequencies in between approximately 20 and 20 000 Hz and sound pressures as low as about -2 dB. The sensitivity of sound pressure is, however, largely frequency dependent; a 4000 Hz tone at 30 dB is not perceived to be as strong as a 100 Hz tone at 30 dB; in fact at 100 Hz, the tone is hardly heard. This dependency needs to be considered when an estimation of a sound's perceived strength is to be made. Equal-loudness contours, shown in figure 2.5, are lines which connect points of equally perceived strength. The lines have been standardised in ISO 226 [25] since 1956 and were first based on [26]. However, the levels of the lines are still subject to research. The standardised levels were changed in 2003; the main changes were made for frequencies below 500 Hz where the lines were raised. An overview of the research can be found in [27].

Another effect that is important to consider is masking, which includes both frequency and temporal masking. In figure 2.6 the frequency masking effect of a sinusoidal signal is schematically shown. It shows the level of a test

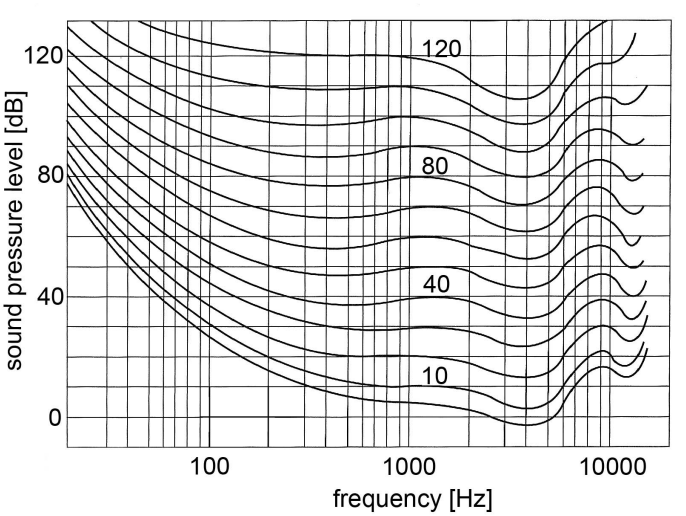

Figure 2.5: Equal-loudness contours of the human ear as reported in [26].

signal, A, that is needed to be audible when another sinusoidal signal, B, is kept constant. The masking effect is larger above the masking signal and the test signal A in figure 2.6 will not be audible. The temporal effect consists of so-called pre- and postmasking effects. A moment after a sound is stopped, and even some time before it starts, it can conceal other surrounding sounds [23].

The last effect addressed here concerns how the auditory system gathers information of a sound's frequency content, i.e. how the auditory filter looks like. As long as a sound signal with a fixed energy have a bandwidth less than a certain bandwidth, called critical bandwidth for loudness, the perceived strength of the sound will be nearly independent of the bandwidth of the sound. The critical bandwidth is thus a measure of the effective bandwidth of the auditory filter. It is often defined empirically; various test methods and tests show similar length of the critical bandwidth. In figure 2.7 the critical bandwidth as a function of frequency is shown [23]. The critical band rate [bark] denotes the critical bands in successive order; there are 24 bands.

2.3 Physical and psychoacoustic measures

Based on the information in the time or frequency domain, single number measures are to be selected or developed to indicate how the sounds are

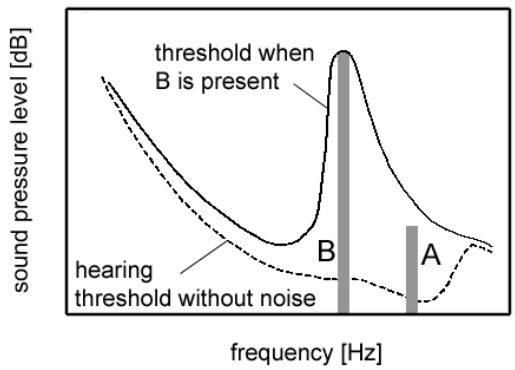

Figure 2.6: Masking by a sinusoidal signal B shifts the original hearing threshold. The test signal A will not be audible.

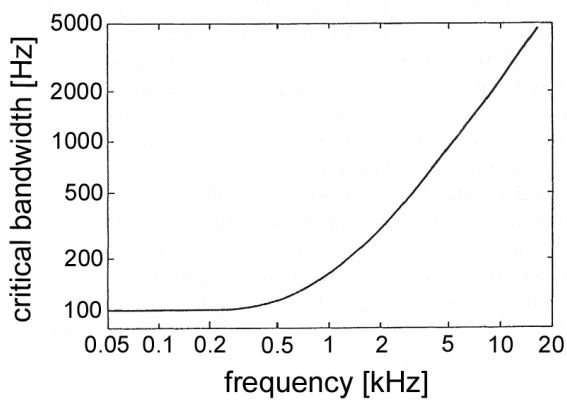

Figure 2.7: Critical bandwidth as a function of frequency [23].

perceived. As has been briefly discussed above, the auditory system is complex and its properties must be considered when selecting single number tools; pure physical measures such as the sound pressure levels in decibels and frequency in hertz are often not enough. Psychoacoustic measures are a means of quantifying sound characteristics in a way that correlates well with human sound perception. Several psychoacoustic measures exist, in the following sections the measures that have been considered in the thesis are presented. For a more elaborate description see [23, 24, 28].

2.3.1 Weighted sound pressure level

Loudness is defined in [24] as "that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud." As loudness is a subjective quantity it cannot be measured directly. However, several models to predict the loudness sensation exists.

One of several methods that has appeared to include the effect in figure 2.5 in a measure of the perceived strength, is the A-weighted sound-pressure level [dB(A)] which was derived approximately from the 40-phon contour. In the same manner, B- and C-weighted sound pressure levels were introduced and derived from the 70- and 90-phon contours[1]. As a result, these measures are applicable to certain levels of the sound. Predictions of the perceived strength of sounds with various frequency and level contents based on, for example, the A-weighted sound pressure level measure are likely to fail.

The great diversity of procedures used to express the strength of sound forced ISO to harmonise the methods. As a first step, the A-weighted sound pressure level measure was standardised, even though its limited applicability was well known. The strength sensation, however, depends not only on the frequency content; other effects, such as masking and sound duration, have to be considered as well.

2.3.2 Loudness

Stevens [29] made several studies to develop a scale of loudness using the method of magnitude estimation described in section 3.2.1. The perceived loudness, L, was suggested to be a power function of a physical intensity, I:

$$L = kI^{0.3}$$

where k is a constant depending on the subject and the units used. Stevens suggested "sone" as the unit of perceived loudness and one sone is defined

[1] There are different information on which contours that was actually used, 40-, 70- and 90-phon contours are those most often mentioned.

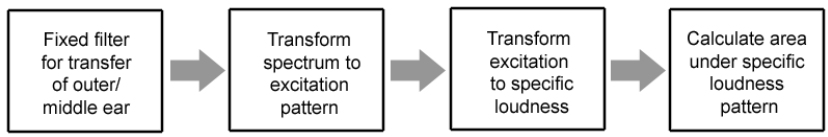

Figure 2.8: Calculation of loudness [24].

as the loudness of a 1000 Hz tone at 40 dB SPL (sound pressure level). The power law model has been confirmed in a large number of studies, although the exponent has seen to be dependent on the nature of the signal [30]. Hence, a more complex model is needed to be applicable on different types of sounds. Three models are mentioned in [24] and were proposed by: Fletcher and Munson [31], Zwicker [32] and Moore and co-workers [33]. The underlying assumption in all these three models are that loudness is related to the total neural activity induced by a sound. Loudness is therefore a sum of the activity in all the critical bands. In figure 2.8 the models' basic structure for the loudness calculation is displayed. In the first step the sound signal is filtered to take into account the transfer through the outer and middle ear. In the next step, the masking effect is included to produce excitation pattern from the spectrum. Thereafter the "specific loudness" is calculated. Specific loudness represents the loudness per critical band and is calculated for each critical band using modifications of Steven's power law. The loudness is thereafter assumed to correspond to the total area under the specific loudness pattern. Loudness is, unlike in the case of A-weighted SPL, a linear measure, i.e., a doubling of the loudness value produces a doubling of the sensation of loudness. This linearity is advantageous specifically in communicating differences and improvements to non-acousticians, and some examples are given in [34, 35].

Loudness is standardised in ISO 532 [36]. Two methods of calculating the measure loudness are described. In part A, loudness is calculated from octave-band analysis, while method B is calculated from one-third-octave band analysis. Part A, proposed by Stevens [37, 38, 39], and part B, proposed by Zwicker [40, 32], do not always agree. Part B, sometimes referred to as Zwicker's loudness, generates generally higher results, but as it is said in the standard, method B seems to take better account of variations in narrow ranges of frequency of the sound spectra. The German standard, DIN 45 631 [41], corresponds to method B. Compared with A-, B- and C-weighted sound pressure level, loudness does not only take account of the level and frequency-dependency of the ear by the use of several equal-loudness contours, but of the effects of masking and spectral distribution as well. Still,

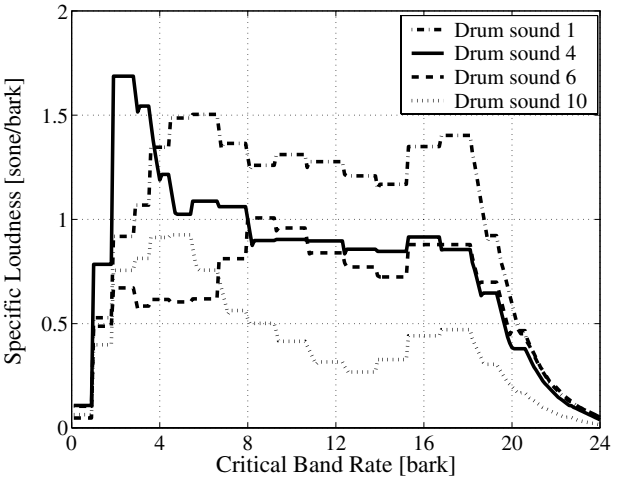

Figure 2.9: Loudness patterns for four drum sounds in [15] where various people wearing different shoes created drum sounds. Drum sound 1 is a 14 mm veneer flooring + PE foam, drum sound 4 is a 10 mm laminate flooring + fibreboard, drum sound 6 is a 7 mm laminate flooring + underlay and drum sound 10 is a 14 mm veneer flooring + polyurethane foam.

a further improvement of the loudness measure should be possible if the latest information on the equal-loudness contours is used in the calculation procedure.

In figure 2.9 the loudness patterns for four of the drum sounds in [15] are shown.[2] Loudness calculations were made using one-third-octave band levels, L_{eq}, for the entire signal (30 s) and thereafter ISO 532B were used. The horizontal axis shows the critical band rate [bark] and the vertical axis the specific loudness [sone/bark]. Drum sound 1 should be perceived as the loudest sound, as the area under its curve is the largest. In [15], the drum sounds were ordered according to their rank of perceived disturbance in the listening test which was shown to also correspond their ranking in loudness. As drum sound 4 has more energy located in the lower region of the critical band rate scale, it should have a darker pitch than the other have. In general, for office spaces, a drum sound with a flat, low curve centred somewhat to the left is sought.

The methods described in ISO 532 are meant for steady sounds and do

[2]The drum sounds are created by male or female gait. It is, however, not the same walking person and the same shoe that are used for all floors, hence the result should not be seen as a guidance of good or less good floor covering solutions for drum sound.

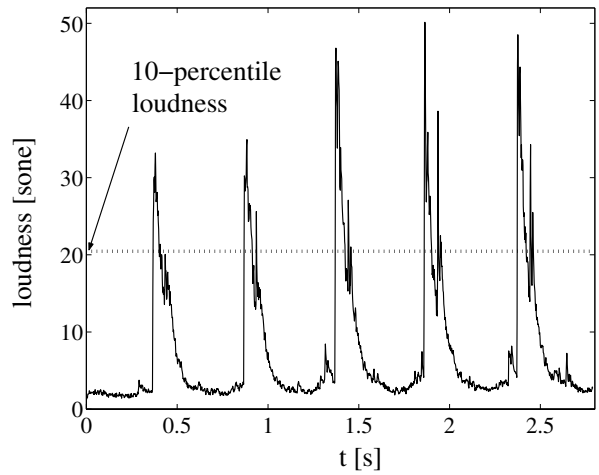

Figure 2.10: Loudness as a function of time from a signal of the same type as in figure 2.2. The figure also shows the 10 percentile, that is the value that is exceeded 10% of the time [15].

not produce a time-variable measure. As the loudness for a tone burst of duration less than 100 ms decreases [23], when measuring on sound with such short duration this behaviour should be included. This effect as well as the effects of postmasking on loudness have been investigated and included in a loudness level meter by Zwicker [42]. Based on the result, loudness as a function of time, loudness percentiles can be calculated. The loudness percentile denotes the loudness that is exceeded in the chosen per cent of time, see figure 2.10. In [15] loudness percentiles were calculated. However, no effects of pre- and post-masking were included and the temporal envelope of the basilar membrane was not represented. Still, the 9-percentile loudness, showed to be the best measure for predicting the perceived disturbance of the drum sounds. It was also shown that even though ISO 532B is meant for steady sounds it is still better than A-weighted sound pressure level. Also in [43, 44] Zwicker's loudness was shown to predict well the perceived loudness of nonstationary sounds.

Although the loudness measure in many cases has shown better correlation to subjective sensations of the sound strength than A-weighted sound pressure level [45, 43, 15], it is not as well known or used, especially not in acoustic standards. Actually, the branch norm [14] might be the first norm for products, where it is used. The still today frequent use of A-weighted SPL is probably due to the fact that A-weighting was introduced in a sound-

level meter in 1936 [46], while the loudness measure was first introduced in a portable level meter in 1981 [47]. The time-consuming calculations were a problem in the past, but modern computers have reduced the effort needed. Increased awareness of the measure among acousticians seems to be needed in order to expand its use further.

An alternative model of loudness, not used in this thesis but still worth mentioning, is the model by Ando [48] which is based on signal analysis methods. Loudness is there calculated from the autocorrelation of the two signals arriving at each ear. Models of pitch, timbre and perception of duration are also presented.

2.3.3 Sharpness

Sharpness is an attribute of timbre. Timbre is defined in [49] as "the aggregate of attributes that allows a listener to distinguish a sound, in terms of subjective impression, from any other sound having the same loudness, pitch and duration as well as the same direction of arrival". In music, different timbre represents the difference between e.g. a violin and a flute when they are playing a note at the same pitch and amplitude. Sharpness is a measure that describes the balance of high and low frequencies in the spectrum of loudness. A low sharpness value indicates high amount of low frequencies and a higher value indicates a higher amount of high frequencies. Its unit is acum. A narrow band noise, one critical band wide, at 1 kHz with a level of 60 dB is assigned to a sharpness of one acum.

Bismarck presented a model of relative sharpness, S/S_0, in [50]:

$$S/S_0 = c\frac{\int_0^{24\text{Bark}} N'\breve{g}(z)\,\mathrm{dz}}{\int_0^{24\text{Bark}} N'\,\mathrm{dz}}$$

where N' is the specific loudness, z is the critical band rate and $\breve{g}(z)$ is a weighting function.

Aurès suggested in [51] a variant:

$$\begin{aligned} S &= 0.08\frac{\int_0^{24\text{Bark}} N'\hat{g}(z)\,\mathrm{dz}}{\ln\left(\frac{N/\text{sone}+20}{20}\right)\text{sone}} \\ \hat{g}(z) &= e^{0.171z/\text{Bark}}. \end{aligned}$$

Zwicker and Fastl's [23] model of sharpness is yet another variant of Bismarck's model and is a weighted centroid of the specific loudness,

$$S = 0.11\frac{\int_0^{24\text{Bark}} N'g(z)z\,\mathrm{dz}}{\int_0^{24\text{Bark}} N'\,\mathrm{dz}}$$

where N' is the specific loudness, z is the critical band rate and $g(z)$ is a weighting function equal to one for critical band rates lower than 16 and has thereafter an exponential growth to reach 4 at the 24^{th} critical band.

The results using these models differ. The model by Aurès is less dependent on the sound level than the others.

2.4 Summary

The drum sound generation was discussed. The human auditory system was described briefly to increase the understanding of the psychoacoustic measures that thereafter were addressed and are used in this thesis.

Chapter 3

Subjective assessment

In the subjective assessment the human response to various sounds is investigated. The focus is here set on assessments where the objective is to improve a product's sound quality, i.e. the suitability of the product's sound. The perception of sound quality is not dependent only on the acoustical signal. In [52] the factors — domains — of sound quality perception are represented as in figure 3.1. The physical and psychoacoustical domains can be described more or less by the sound pressure level, the temporal, frequency, and spatial distribution, loudness, sharpness, etc. The design domain can be divided into three fields where the attention is on working with:

(1) stimulus-response compatibility, where the relation between a stimulus and its response is investigated. As an example, a sound can be designed to carry information of the functioning of a machine.

(2) pleasantness of sound, where the evaluation concerns the pleasantness/unpleasantness and similar terms.

(3) identifiability of sounds, where studies on, for example, speech or the right product sound are performed [53].

The cognitive domain can be divided into three groups:

(1) source-related, the image of the source, i.e. a sport car or a family car

(2) situation-related, e.g. is it the person sitting on the motorbike accelerating or the person sitting in his garden nearby that assesses the sound

(3) person-related factors, e.g. expectation, motivation, preference, etc. [4]

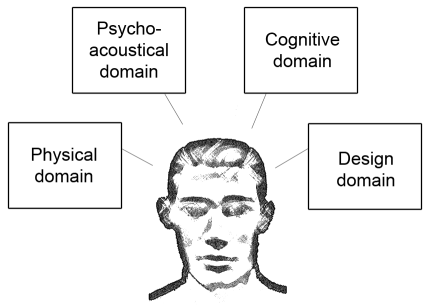

Figure 3.1: Domains of sound quality perception [52].

Sound quality assessments are usually performed in laboratory. Naturally, the laboratory result can differ from a field-test where the sound is produced in a normal-life situation. A field test has the following advantages [54]:

- it creates a representative situation of the sound in daily life;
- no recordings of the sound are needed — the recordings are time-consuming and even though advanced recording techniques exist it is very hard to reproduce the sound so that the human hear cannot distinguish between original and reproduced sound;
- if a product is evaluated, a typical handling of and interaction with the product is enabled;
- assessors can individually select a typical or critical situation or state to base their opinion upon.

On the other hand, laboratory tests have following advantages:

- the test is reproducible;
- all assessors have identical test conditions;
- if products are compared, identical states of operation can be presented;
- the sounds can be compared directly, decreasing the influence of loss of sound memory;
- modifications of sound can be made before and during the test;
- the test is time-efficient.

Laboratory tests must be planned so that the results can be transferred to the field. Influencing factors should be identified. If factors that influence the response cannot be varied in a laboratory, they should at least be kept constant during the test. The most difficult factors to present correctly are the cognitive factors, but with the help of additional information such as verbal descriptions, photos, etc. the laboratory environment's deficiencies can be decreased. Whenever test results are presented, it is necessary to ensure that the information given in the test is enclosed.

Drum sound evaluations can be made by letting the assessor walk on the test floors. However, one must consider that the subjective response might differ when the drum sound is created by oneself compared to when it is created by someone else. A disadvantage of letting people walk and assess the drum sound is that other criteria, such as the springiness of the floor, can cause an unconscious effect on the assessor. In office spaces, it is mainly the drum sound created by people other than the listener that is a problem. Therefore, if the objective of the investigation is to improve the drum sound in office spaces, the evaluation should be made by letting the assessor listen either to other people creating drum sound live or to recorded drum sound. The listening tests in this thesis were all performed in an office environment using recorded drum sound. As the test situation was the same as the situation the assessors were asked to imagine themselves in, any bias due to cognitive factors is reduced.

References to various investigations on sound quality using different approaches can be found in [53]; in the same volume of Acustica united with acta acustica valuable articles concerning sound quality can be found. Another useful reference, providing a guideline to performing listening tests, is [55]. International standards for performing listening tests exist for some areas, for example noise annoyance [56] and loudspeakers [57].

3.1 Recording and reproduction

An important part of performing tests in laboratory is making adequate recordings and reproductions of the sound. The are several ways to pick up the sound signal. In mono channel measurements, one microphone is used. The time and frequency content are given but no directional or localisational information. Cues across one axis are given by stereo recording which uses two different channels of the audio signal, either recorded with two microphones spaced apart or with a single microphone with two elements. "Dual mono" means two channels coming from the same microphone. The basic idea of the binaural recording technique is to record and reproduce signals

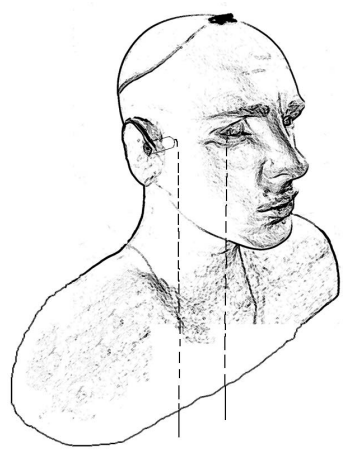

Figure 3.2: The dummy head used at the division of Engineering Acoustics. Two microphones are located at the entrance of the ear canal.

to human hearing accurately. Three axes of localisation cues are provided. Its history goes back to 1881, and the first binaural unit was an array of carbon telephone microphones installed in the Paris Opera. The signal was sent through the telephone system, and required that the subscribers wear a special headset, which had a small speaker for each ear. Two omnidirectional microphones are placed either at the ear canal entrance or at the eardrum even though measurements at the entrance of the ear canal are recommended [58]. The use of an artificial head, also called a dummy head (see figure 3.2), is prevalent even though it is argued that the use of microphones in the entrances of a real person's ear canals can be at least as effective as using a dummy head [59]. (In some cases, such as in small cabins of work machines, the use of a dummy head is not possible while operating [60].)

Binaural sounds can also be achieved by the use of Head-Related-Transfer-Functions (HRTF). The transfer functions from a location of a sound source to the left and right ear are then measured enabling synthesised accurate binaural signals from a monaural source. Each individual has its own HRTF; a dummy head has some kind of mean HRTF's. Therefore, when listening to the recordings using a dummy head compared with the original sound, some deviations may occur. Localisation errors are common. So even though it was stated above that binaural technology provides three axes of localisation cues, the true location of the actual sound source might not be correctly reproduced. Still, binaural recordings with playback through headphones generally give the most natural sound reproduction. Sounds with very low

frequencies that are felt by the whole body, an additional sub-bass loudspeaker might be needed. A review of the binaural technology and various techniques to record and reproduce sound can be found in [4, 61].

3.2 Evaluation methods

Various methods to perform listening tests are available; their applicability is dependent on the aim of the study and its prerequisites. In a unidimensional analysis one dimension of the sound is investigated. In a multidimensional analysis, several dimensions and how they interact are investigated. Multidimensional analysis can be valuable, especially at the start of a research investigation, for learning more about the relevant dimensions. Even though the methods are presented here in the section on unidimensional analysis, their result can, in a following analysis, be used in a multidimensional analysis. The methods presented are mainly for sound quality tests. [23] contains additional methods to be used in psychoacoustic investigations such as threshold measurements. For every method there are pitfalls to avoid such as stimulus range and sequence effects; see Poulton [62] for an overview.

3.2.1 Unidimensional analysis

The response is usually a composite of various things. A drawback of unidimensional analysis is that slight variations in the design of the question can produce large variations in response. However, techniques to decrease this risk have been developed and are applicable in sound quality tests as well. Several analysis tools have been developed for the study of attitudes. Thurstone is the social psychologist who first created attitude-measurement methodology. He developed three methods: the method of paired comparisons, the method of equal-appearing intervals and the method of successive intervals. All three methods are based on his *law of comparative judgement*. It assumes that every given stimulus, T_i, is associated with a sensation, X_i, and that this process can be ordered on a psychological scale or continuum. How a group of stimuli is ordered depends on the attribute that is of interest. However, even though the attribute is set, a given stimulus does not always produce the same sensation to an assessor but fluctuates around its "true" scale value, S_i = mean of X_i. The scale is then defined so that the fluctuations form a normal distribution. Thurstone's methods will be described here briefly as will the method of magnitude estimation and Likert scales. The section on methods of paired comparison is not restricted to Thurstone's model. For a more extensive description of Thurstone's models see [63].

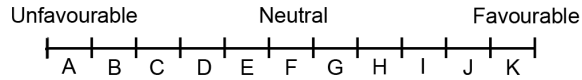

Figure 3.3: Thurstone equal-appearing interval scale.

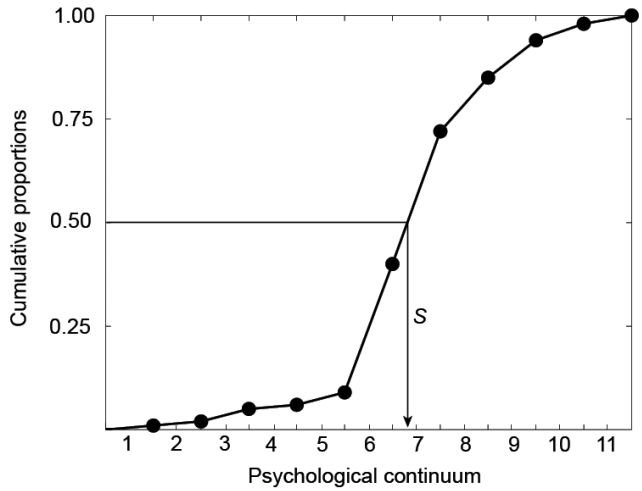

Figure 3.4: Cumulative proportion graph showing how the scale value, S , is determined.

Method of equal-appearing intervals

In the model's initial form each assessor (judge, listener) is asked to sort statements (stimuli) into 11 categories named A–K so that the intervals between categories are subjectively equal. The A category represents the least unfavourable whereas K is the most favourable. F is the neutral category, see figure 3.3. The whole range of stimuli is presented prior to the test and the assessors are instructed to use the full range of the scale.

If the intervals are believed to have been judged equal by the assessors, the categories are assigned numbers from 1 to 11. The number of times a stimulus was placed in each category, here called the frequency, gives the proportion of judgements, that is, the frequency divided by the total number of assessors. The cumulative proportions are thereafter calculated. The scale value of the stimulus is then given at the point where the cumulative proportion is equal to 0.50, see figure 3.4.

Numerous variations of this model exist using various numbers of categories and descriptions [63, 53]. When creating a scale, consider the number

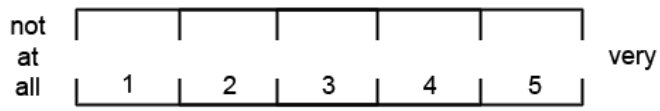

Figure 3.5: Five-point response scale with verbal endpoints.

of items included in it. Many items on a scale can make the scale more reliable and reduce the risk that the score is due to error. On the other hand, too many items can create a problem. Multiple items can, unintentionally, focus on different aspects of the question and may, therefore, not be homogeneous. The scale will then no longer be measuring just one question, but several! The risk of verbal scales is that they might not have equal distances between the labels and therefore parametric analyses cannot be applied. An example of a verbal scale is the Likert scale [63] where five categories are commonly used: strongly agree, agree, neutral, disagree, and strongly disagree. One cannot always assume that the assessor means that the difference between agreeing and strongly agreeing is the same as between agreeing and being neutral, and therefore parametric studies such as arithmetic mean should not be calculated.

The scales can be either unipolar in the sense that they describe the intensity of an attribute (e.g. not loud – loud), or bipolar so that the words at the ends describe opposite attributes (e.g. dark – bright). For bipolar scales there are discussions about whether they should have an even or odd number of categories. An even number forces the assessor to choose between higher and lower scores, which might look attractive but it might also hide information on the difficulty of the test. In [53] the scale in figure 3.5 is recommended. It is advantageous as it can be applied to several attributes. Moreover, the difficulties of using a bipolar scale to find contrasting adjectives that truly belong to the same dimension are avoided.

Method of successive intervals

This is an extension to the equal-appearing intervals scaling. The instructions to the assessor are the same and the cumulative proportions are calculated in the same manner. The width of each category is, however, not set to one but is estimated for each category. It is then assumed that the cumulative proportions are normal when projected on the psychological continuum.[1]

[1]This assumption can be checked by plotting the cumulative proportion distribution for the stimuli in a normal probability plot [63].

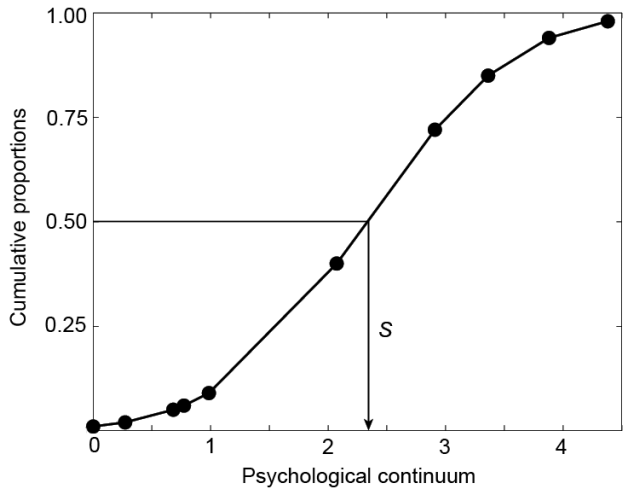

Figure 3.6: Cumulative proportion graph showing how the scale value, S , is determined. Note that the widths between each category as represented by a dot, are not equal as in figure 3.4.

The difference in successive intervals' normal deviates is calculated for each stimulus in the test. The width of each category is estimated by the mean width of all stimuli. In figure 3.6 a cumulative proportion graph is shown. Note that the width between each category is not equal as in figure 3.4. When the result looks like figure 3.6, it would appear that even though the assessor was told that the categories were of equal size, they were not used in that way.

Method of magnitude estimation

Stevens [29] developed the technique to relate perceived magnitude and stimulus intensity. In the method of magnitude estimation it is assumed that the assessors (listeners) in a listening test can assign numbers to the perceived attribute of the test. A standard sound can be presented with an assigned value (say 100) to which all subsequent stimuli are rated relative, either with a difference in number or with a ratio. This method is debated; the responses using ratios and differences have sometimes given different results. Poulton [62] concludes that only when familiar units are used should they be related by arithmetic rules, otherwise bias occurs.

An example of a study where the method of magnitude estimation is used is a study on the effect of a preference for rock music on magnitude

estimation scaling behaviour in young adults [64]. It is shown that the assessors who disliked rock music provided significantly higher mean numerical responses for all intensities presented. It was concluded that they perceived the stimulus as being louder simply because they did not like rock music. Listening tests are sometimes divided into objective (perceptive - examining what persons hear) and subjective (affective – examining what persons prefer or dislike) tests [55]. However, the study above is an example of how elusive this division might be; even though the assessment initially might have been considered objective, the result shows that it is not.

An ISO-standard developed for food products concerning magnitude estimation can be found in [65].

Paired comparison methods

Scales have an advantage in their ability to produce an absolute value corresponding to a certain sensation. However, scales can be hard for an assessor to use — uncertainties about whether the assessors have used and understood the scale equally can occur — and it can therefore be difficult for the researcher to analyse the result. A way of avoiding this problem is to use paired comparison tests. In such a test assessors are asked to tell which of two treatments — or stimuli or sounds — has a certain attribute (such as a pleasant sound), see figure 3.7. The need for the assessor to have a longer sound memory is eliminated and more consistent answers can therefore be achieved. In [66] a method of paired comparison, method of equal-appearing intervals and method of successive intervals were compared in the evaluation of annoyance response to engine sounds. It was concluded that consistent judgements of annoyance were observed with the paired comparison data. Judgements of annoyance using the methods of equal-appearing intervals and successive intervals were made consistently only by trained assessors.

One disadvantage of paired comparison models is that no absolute values corresponding to a certain attribute are given, but only relative values. Another disadvantage is the rapid growth of comparisons needed when the number of treatments increases and all pairs are compared. This could, however, be treated to some extent by the use of incomplete balanced design where not all pairs are compared in the test. The design is explained in [67] and tables that enable balanced incomplete designs can be found in [68, 69].

Two major basic pair comparison methods exist today: the Thurstone-Mosteller [70, 71] model and the Bradley-Terry model [72]. These models are both so-called linear models. The linear paired comparison model is defined by David [67] as

$$P(T_i \rightarrow T_j) = H(V_i - V_j)$$

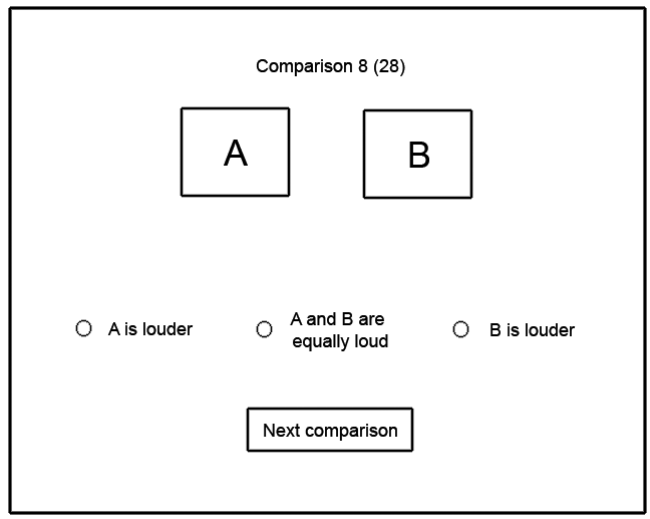

Figure 3.7: Example of layout in paired comparison tests allowing ties as used in paper C [16].

that is, the probability of choosing treatment T_i when compared with T_j is a function of the difference in their strengths (or scores) V_i and V_j only, where the function H is a symmetric cumulative distribution function, $H(-x) = 1 - H(x)$. The Bradley-Terry model assumes a standard logistic distribution function and the Thurstone-Mosteller model assumes a normal (Gaussian) distribution function.

Sometimes the assessors are not able to reveal any difference in the pair. This is the case when the treatments are equal in the specific attribute assessed or when the difference is too small to be perceived. When ties are not allowed, the assessors are forced to make a selection and the choice is, hopefully, made randomly. When no models for handling ties were developed, a way of excluding this random behaviour from the input data was to allow ties in the test but to ignore them in the analysis. Even though models that do allow ties in the analysis were developed during the 60's and 70's [73, 74, 75], there are still situations where they are not used today although they could increase the information in the analysis.

In paper A [5], aspects on these methods are given. It is seen that the choice of either Thurstone-Mosteller or Bradley-Terry is not crucial. As the former model provides an algebraic solution, it is recommended when scale values are requested. However, when estimations of their differences are to be made, when the treatments are inhomogeneous or if the design is either

incomplete or unbalanced, the latter model is recommended. When preference is the objective, ties should be allowed as they add information and can decrease the results' confidence intervals/regions. The choice between the Rao-Kupper [74] and Davidson [75] models in terms of their ability to handle ties is not critical.

Several extensions and variations to the paired comparison model exist; for example in [76] a method to include effects of time-varying data is presented. In cases where more than a three-point scale ($T_i \rightarrow T_j$, $T_i = T_j$, $T_i \leftarrow T_j$) is requested the method by Scheffé is recommended [77]. It does not only provide a means of analysing data based on a 7- or 9-point scale but within-pair order effects can also be investigated. Many aspects that are not dealt with in paper A, such as within-pair order effects, circular triads, consistency tests, triple comparisons and multivariate paired comparisons where several characteristics in the treatments are investigated, are addressed in [67].

3.2.2 Multidimensional analysis

In the following part of this section approaches to performing multidimensional questionnaires, to be used with or without additional collected data as instrumental measurement results later on, are presented. In section 4 multivariate methods, where the relation between several characteristics in the treatments are investigated, will be discussed.

Multiple unidimensional scales

By combining several attributes of a sound using any unidimensional scale one after the other as in figure 3.8, a following multidimensional analysis is made possible. An advantage of multiple unidimensional scales is that several scales using different attributes for describing the same dimension can provide data to perform reliability tests and control to ensure that the decision is not based on extraneous criteria that might be introducing bias into the test. The gain might, on the other hand, be offset by the possible effect of several scales confusing the assessor. The adjectives in figure 3.8 could be used when evaluating drum sound, however, the word "natural" indicates that the assessor is informed of the included type of floor covering as a natural sound is different for a stone or wood floor. When selecting adjectives for a test where there is little knowledge of which adjectives are important, the use of several adjectives in the initial tests is crucial. By the use of some kind of factor analysis (see section 4.2) the dimensions and

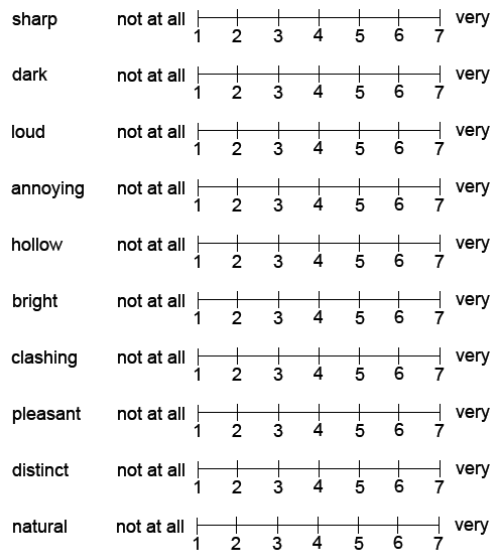

Figure 3.8: Example of multidimensional scales.

adjectives can be decreased in the following tests. Studies where multiple unidimensional scales are used can be found in [78, 79, 80].

Semantic differential

The semantic differential is an extension to the unidimensional scale. It was first developed by Osgood [81, 82] as a tool for research on the psychology of meaning where the connotative meanings of certain words were examined. He constructed bipolar scales based on semantic opposites, such as "good–bad", "beautiful–ugly", etc. The scales were called "semantic differential" scales because they differentiated the attitude's intensity based on how a person interprets the connotative meanings of words. The approach has been used in several sound quality investigations; it was probably first used in the late 50's by Solomon [83], who examined sonar sound.

The scale usually consists of 7 possible grades. The semantic scales need to be adjusted to fit the topic of research. When presenting the scales, make sure the variables that might be expected to affect the assessment are not presented always on one of the sides as it might introduce bias. An advantage of using semantic differentials is that fewer scales can be used compared with the method using adjectives and "not at all—very"; one scale for example

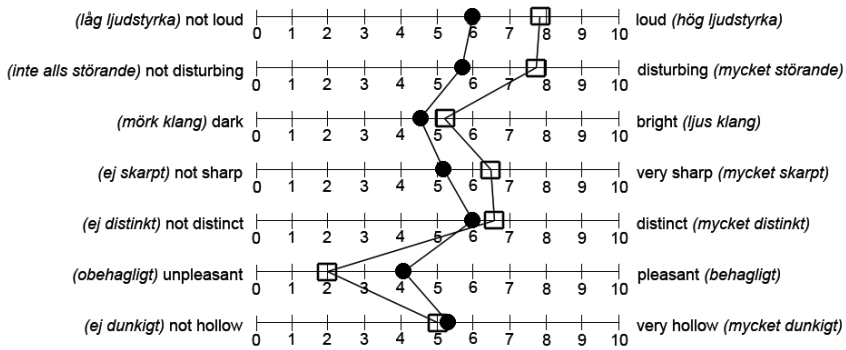

Figure 3.9: Polarity profile of two drum sounds in [6]. Filled circles: laminate flooring + fibreboard; open squares: 14 mm veneer + PE foam. The Swedish words used in the test are shown in italics.

"dark—bright" can be used instead of "dark: not at all—very" and "bright: not at all—very", hence fewer judgement are needed. However, as discussed earlier for bipolar scales, it is important that the words belong to the same dimension. The construction of such scales needs careful consideration; for example what is the antonym of "sharp" that was used in figure 3.8? If "pleasant" is used, a bias is introduced since it is assumed that a sharp sound must necessarily be unpleasant. The problem can be solved by adding a "not" in front of the adjective, for example, sharp–not sharp. (Solomon [83] used sharp–dull). In the literature, a mixture of semantic opposites and adjectives with and without their negatives are often used (car interior [84], household appliances [85], hair dryer [86]).

In a semantic polarity profile the scalings of each stimulus are combined. In figure 3.9 the polarity profile for a drum sound test in [6] is shown. Comparison between different stimuli is then enabled. The scalings can also be used in a factor analysis or cluster analysis to determine if the number of connotative dimensions can be reduced.

Similarity ratings for pairs of sounds

In this method the assessor is asked to scale the degree of similarity or dissimilarity between pairs of sounds. As the similarity or dissimilarity of every pair is requested, the method becomes cumbersome for large numbers of stimuli. The results are examined by statistical multidimensional scaling (MDS) analysis; see section 4.3 where the dimensions that describe the similarities/dissimilarities are sought. The disadvantage of the method, which

also might be an advantage, is that the meaning of the dimensions that are found can only be described by interpreting the physical properties of the sounds that are located near each other.

3.3 Summary

Factors influencing the perception of sound are discussed as are also advantages and disadvantages of performing listening tests in field or laboratory. Various evaluation methods, both unidimensional and multidimensional, are reviewed. Scales have an advantage in their ability to produce an absolute value corresponding to a certain sensation. However, bias can be introduced into the test unless the construction of the scales is carefully considered. Paired comparisons are easier to perform for an assessor and give, in general, more consistent responses. The disadvantage of paired comparison is the rapid growth of comparisons needed when the number of treatments increases and all pairs are compared. Another disadvantage is that no absolute values corresponding to a certain attribute are given, but only relative values. In this thesis, the method of paired comparison is used in an office environment using recordings of the drum sounds.

Chapter 4

Relating subjective and objective assessment

With the information from the listening test and the objective measures, the search for any relations between them can begin. Some kind of multivariate analysis is often used. It is a statistical analysis technique in which multiple variables are analysed separately to determine the contribution made by each variable to an observed result. Examples of multivariate analysis techniques are multiple regression analysis, factor analysis (including principal component analysis) and multidimensional scaling; these will be described briefly in the following sections. There are numerous other multivariate analysis techniques; the ones described here were chosen as they are common in sound quality assessments. For a more elaborate description, consider the statistical literature on multivariate analysis [87]. A nice overview can be found on the internet [88].

These methods will differ from the methods used traditionally in building acoustics. By the use of various reference curves to which the instrumentally measured sound in decibels is compared, the insulation of walls and floor structures is evaluated to adjust the physical measurement to the performance in subjective listening tests [89, 90]. Here, a combination of acoustic and psychoacoustic measures is used to find any relation between subjective and objective assessment. However, the result using the reference curves can be used as an input in the multivariate analysis.

4.1 Multiple regression analysis

Multiple regression is the simplest of the multivariate analysis techniques. This method was used in paper B [15] to find a prediction model of the

subjective response to drum sound. The multiple linear regression model describes the relationship between a single dependent variable or response, y, and n independent or regressor variables, $x_1, x_2, ..., x_n$, as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_n x_n + \epsilon$$

where $\beta_j, j = 0, 1, ..., n$ are called the regression coefficients or partial regression coefficients. β_j measures the expected change in y per unit change in x_j when all the remaining independent variables are held constant. ϵ is the random error. The term linear is used as y is a linear function of the unknown parameters $\beta_j, j = 0, 1, ..., n$. Although they are called independent variables, in multiple regression the independent variables may be correlated to some extent. However, if the independent variables are highly correlated, or when the number of independent variables is large compared with the number of observations, multiple linear regression is not applicable and partial least square regression should be used instead, see [91].

Hence, in multiple regression, there is one dependent variable and several independent variables. In simple regression, there is only one independent variable; in factor analysis and most other multivariate techniques, there are several dependent variables. Multiple regression analysis can be extended to handle several dependent variables and is then called multivariate regression analysis [87, 92]. However, it will not be discussed further here.

The model fitting is typically made using the method of least squares where the β's are chosen so that the sum of the squares of the errors is minimised.

The significance of the model can be tested in a hypothesis test. It is then tested to see whether at least one of the β's is contributing significantly to the model. The test is done using analysis of variance (ANOVA) techniques. The coefficient of determination, R^2, is a measure of the amount of reduction in the variability of y that is due to the independent variables $x_1, x_2, ..., x_n$, $0 \leq R^2 \leq 1$. However, a large R^2 does not automatically indicate that the regression model is a good one; when a variable is added to the model, R^2 is always increasing, independent of whether it is a statistically significant variable or not. With the use of the adjusted R^2 statistic this behaviour is decreased. When R^2 and adjusted R^2 differ dramatically, nonsignificant variables are probably added [93].

4.2 Factor analysis

The objective of factor analysis is to discover simple patterns in the relationship pattern among the variables. More precise, it seeks to discover if the

observed variables can be expressed in terms of a smaller number of variables called factors. Different methods to extract these factors from a set of data exist [87]:

- Principal components analysis, PCA: The most common form of factor analysis. PCA seeks a linear combination of variables (principal component) that accounts for as much of the variability in the data as possible. This variance is then removed and a second linear combination is sought which explains the maximum proportion of the remaining variance, and so on. The factors are orthogonal (uncorrelated). PCA analyses the total (common and unique) variance.

- Principal factor analysis, PFA: Also called principal axis factoring and common factor analysis. PFA seeks the lowest number of factors which can account for the common variance of a set of variables.

- Other extraction methods: alpha factoring, image factoring and maximum likelihood factoring, unweighted least squares factoring, and generalised or weighted least squares factoring.

PCA should be used when approximating the data using fewer dimensions, PFA when an explanatory model for the correlations in the data is sought [94]. Only PCA will be described here as it is dominant in the acoustic papers.

Principal component analysis partitions the total variance, i.e. the sum of the variances in the original data, by first finding the linear combination of the variables, $x_1, x_2, ..., x_n$, that accounts for as much of the variability in the data as possible.

$$y_1 = e_{11}x_1 + e_{12}x_2 + ... + e_{1n}x_n$$

y_1 is called the first principal component. The procedure is displayed graphically in figure 4.1. After having removed the variance in the original data that can be attributed to the first component, the second component accounting for the next largest amount of variance is sought. This component is constructed to be uncorrelated (orthogonal) to the first component. The procedure is repeated until the variance still to be explained is small enough. Then how are the coefficients $e_{ij}, i, j = 1, ..., n$ found? Suppose the original data is given in vectors $\mathbf{x}_1, ..., \mathbf{x}_n$. Estimates of the sample mean $\mu_\mathbf{x}$ and covariance matrix $\mathbf{C}_\mathbf{x}$ (also called dispersion matrix) are

$$\begin{aligned} \mu_\mathbf{x} &= E\{\mathbf{x}\} \\ \mathbf{C}_\mathbf{x} &= E\{(\mathbf{x} - \mu_\mathbf{x})(\mathbf{x} - \mu_\mathbf{x})^T\} \end{aligned}$$

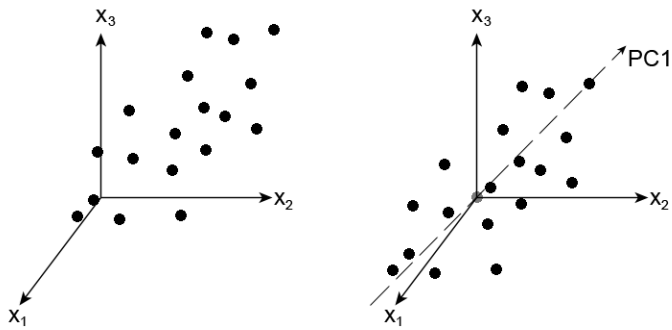

Figure 4.1: Left: Every variable represents one co-ordinate axis; for simplicity only three variables are shown. The length of each co-ordinate is standardised usually to unit variance scaling. The observations are placed in the n-dimensional space. Right: The procedure involves subtracting the averages of the data which correspond to a re-positioning of the co-ordinate system so that the average point (grey point) is the origin. The first principal component, PC1, is the line that best fits the data in the least squares sense. Every observation is thereafter projected to this line and its co-ordinate on the PC1 line is its *score*.

where E denotes the expected value operator. From the symmetric covariance vector an orthogonal basis can be calculated by finding its eigenvalues and eigenvectors. The eigenvectors $\mathbf{e}_i$ and corresponding eigenvalues λ_i are the solution of

$$\mathbf{C}_x\mathbf{e}_i = \lambda_i\mathbf{e}_i, \quad i = 1, ..., n$$

where the eigenvalues are found solving

$$|\mathbf{C}_x - \lambda\mathbf{I}| = 0$$

$\mathbf{I}$ is the identity matrix with the same order as $\mathbf{C}_x$. The eigenvector of the largest eigenvalue corresponds to the direction of largest variance and hence gives the coefficients e_{ij}.

As shown in 4.1 the projection of each observation on the principal component line gives its *score*. The scores are often displayed in a score plot. In figure 4.2 a PCA score plot of the first two PCs of drum sound data from [21] is displayed. The results from the paired comparison listening test and the objective test is used. The meaning of the scores can be given by the *loadings* in a loading plot, see 4.3. The variables' loadings for each principal component are given by the coefficients e_{ij} and the standard deviation of that is accounted for by the principal component. Strong influence of a variable to an observation's score is achieved when location of the variable's

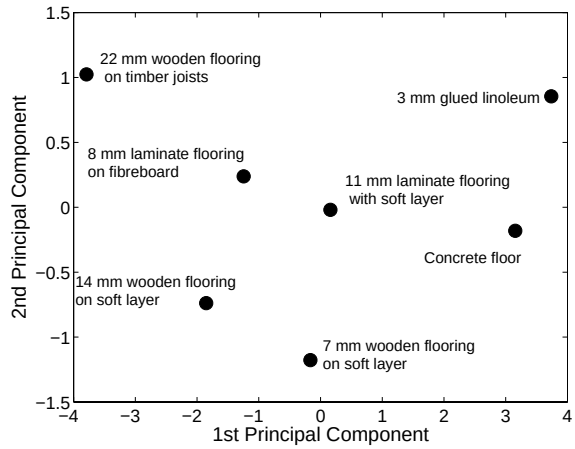

Figure 4.2: PCA score plot of the first two PCs of drum sound data from [21].

loading corresponds with the location of the observation. The distance to the origin conveys information on how strong an impact the variable has on the component; the further away, the stronger the impact. In the score and loading plot it is possible to group observations (e.g. drum sounds) with similar attributes and variables that contribute with similar information. In figure 4.3 it is seen that the first component, which explains more than 90% of the variance, consists of measures of the sound's amplitude (strength); the second, the influence of which is much less, about 8%, consists mainly of the sharpness measure. As the perceived loudness and disturbance are located at the same position as the amplitude measures, it is likely that the amplitude measures have a strong influence on the perceived loudness and disturbance. The small influence of sharpness corresponds with what was seen in [21], where sharpness did not improve the correlation between the perceived loudness/disturbance and the objective amplitude measures.

In sound quality investigations PCA is often applied to find how many dimensions account for most of the variance of the result on multiple unidimensional scales or semantic scales. PCA is a common tool, especially in studies of the perceptual evaluation of sound reproduction systems [78, 95, 96]. In [97, 98] it is used for sound quality evaluation of diesel engines and in [80] to define the sound character in car compartments. The result from a PCA can also be used to find variables to be used in a following multiple regression. Moreover, as the new components are uncorrelated, any problems in multiple regression that might have occurred if the original data consisted of correlated parameters are removed by the use of the PCA components [87].

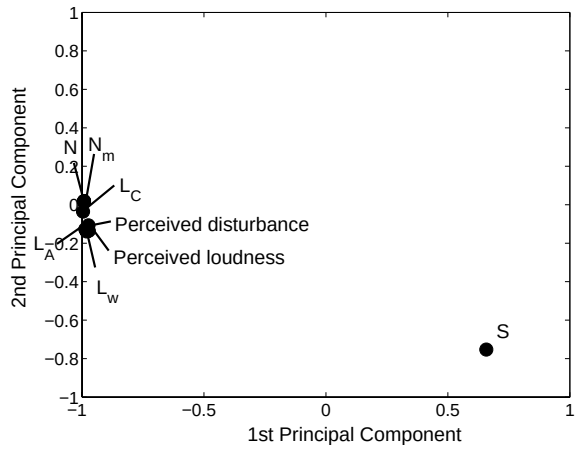

Figure 4.3: PCA loading plot of the first two PCs of data from [21]. Perceived loudness and disturbance from paired comparison listening test; the objective instrumental measures: N_m is the loudness measure according to [14], N is the loudness according to [36], L_A and L_C are the A- and C-weighted sound pressure levels, L_w is the weighted sound pressure level according to EN ISO 717-2 [89], clause 4, and S is the sharpness according to Zwicker and Fastl [23].

However, an increasingly common regression tool is the partial regression technique that generalises and combines features from PCA and multiple linear regression, see [91] for more information.

4.3 Multidimensional scaling

Multidimensional scales, MDS, are nicely introduced in [99] as follows:

> "Suppose a set of n objects is under consideration and between each pair of objects (r, s) there is a measurement δ_{rs}, of the 'dissimilarity' between the two objects. For example the set of objects might be ten bottles of whisky, each one from a different distillery. The dissimilarity, S, might be an integer score between zero and ten given to the comparison of the r^{th} and s^{th} whiskies by an expert judge of malt whisky. The judge would be given a tot from the r^{th} bottle and one from the s^{th} and then score the comparison: 0 – the whiskies are so alike she cannot tell the difference, to 10 – the whiskies are totally different. The judge is presented with all forty-five possible pairs of whiskies, and after

> a pleasant day's work, provides the data analyst with a total set of dissimilarities δ_{rs}. $\cdots$ A narrow definition of multidimensional scaling (often abbreviated to MDS) is the search for a low dimensional space, usually Euclidean, in which points in the space represent the objects (whiskies), one point representing one object, and such that the distances between the points in the space match, as well as possible, the original dissimilarities δ_{rs}. The techniques used for the search for the space and the associated configuration of points form metric and nonmetric multidimensional scaling."

In the above example, the expert judged dissimilarities of pairs using a one-dimensional scale. In a subsequent analysis other input data given in vectors, $\mathbf{x}_1, ..., \mathbf{x}_n$, that were collected or measured in other ways are combined with the judge's response creating a multidimensional space to find out the parameters that influenced the judge's response. MDS can be considered as an alternative to factor analysis. In factor analysis the similarities are expressed in the covariance matrix; in MDS any kind of similarity or dissimilarity matrix can be used. MDS refers to a group of methods with slightly different optimization algorithms to find the sought parameters. The first well-known MDS proposal was made by Torgerson [100]. One least square method to find the influencing parameters is to minimize the stress measure, ϕ:

$$\phi = \sum_{r \neq s} \left[d(r,s) - d'(r,s) \right]^2$$

where the distance between r and s is denoted by $d(r,s)$ and the distance between r and s in the estimated space is denoted by $d'(r,s)$.

As already mentioned, there are numerous methods belonging to MDS, to find more information on Multidimensional scaling see [99]. MDS has been used for sound quality assessment in [101] and in [102]. In [102], MDS is used as one step out of four to find how the sound quality of a sound is perceived. The four steps are (1) Semantic scale evaluation using PCA, (2) Multidimensional scaling, (3) Preference mapping using paired comparison tests and (4) Synthesis of the results. The use of various methods in different steps where each methods' advantages can be fully utilized, is probably a good approach to learn more about the character of complex sounds and to improve sound quality.

4.4 Summary

Acoustic and psychoacoustic measures are used to find any relation between subjective and objective assessment. Multiple regression analysis, principal component analysis, PCA, and multidimensional scales, MDS, are presented. It is mainly the multiple regression analysis that is used in this thesis. Multiple regression are particularly useful when there is some knowledge of the influencing parameters, PCA is advantageous when many variables are included and the number of influencing dimensions is sought. The advantage of MDS is the simple test situation for the assessor as only the level of differences or similarities is assessed; the disadvantage of MDS is that the meaning of the dimensions that are found can only be described by interpreting the physical properties of the sounds that are located near each other. The use of various methods in different steps where each methods' advantages can be fully utilized, is probably a good approach to learn more about the character of complex sounds and to improve sound quality.

Chapter 5

Future Work

Many parts of the drum sound field still need to be studied. Some topics are listed here:

- A measure of the drum sound's timbre is needed to further describe the character of the drum sound.
- The applicability of the findings on floorings built up on joists, or raised/access floorings needs to be further examined.
- The parameters in figure 1.3 that are not addressed in this thesis need to be investigated, that is, the influence of walking speed, background noise and room acoustics.
- A theoretical model of the interaction of the foot and floor system and the produced sound radiation is needed to enable further improvement of the drum sound.
- Drum sound measurements in field (both subjective and objective) to find out the proposed drum sound norm's applicability in field.

Bibliography

[1] F. Larris: Drum noise from floors. Teknologisk Institut, Lydteknisk Konsultation, København, In Danish, 1952.

[2] O. Brandt: Akustisk planering, In Swedish. Stockholm, 1958.

[3] U. Jekosch, J. Blauert: A semiotic approach toward product sound quality. Proceedings of InterNoise 96, GB-Liverpool, 1996. 2283–2286.

[4] M. Bodden: Instrumentation for sound quality evaluation. Acta Acustica united with Acustica **83** (1997) 775–783.

[5] A.-C. Johansson, E. Nilsson, P. Hammer: Aspects on paired comparison models for listening tests. Submitted to Acta Acustica united with Acustica, 2005.

[6] E. Nilsson, P. Hammer: Fotstegsljud från olika typer av golvbeläggningar (In Swedish). Technical report TVBA–3104, Engineering Acoustics, LTH, Lund, Sweden, 1999.

[7] A.-C. Johansson, E. Nilsson, P. Hammer: Footstep sound from different floor coverings, subjective measurements. Acoustics 2000, Volume 22, GB-Liverpool, 2000. 95–100.

[8] NF S 31–074, Acoustics – Measurement of sound insulation in buildings and of building elements – Laboratory measurement of in-room impact noise by floor covering put in this room. Norme francaise, AFNOR, 2002.

[9] D. Hoffmeyer: Measurement of drum noise – A pilot project. Tech. Rept. NT 1597-02, Nordtest, Finland, 2004.

[10] B. Plinke: Optimierung des akustischen Verhaltens von Laminatfussböden. Techn. Rep. 12207N, Fraunhofer-Institut für Holzforschung, Wilhelm-Klauditz-Institut (WKI), Braunschweig, 2002.

[11] EN ISO 140-8:1997, Acoustics – Measurement of sound insulation in buildings and of building elements – Part 8: Laboratory measurements of the reduction of transmitted impact noise by floor coverings on a heavyweight standard floor.

[12] B. Plinke, J. Gunschera: Harmonisierung in Sicht : Prüfverfahren für das Raumschallverhalten von Laminatböden. Laminat-Magazin, Januar (2002) 18–24.

[13] E. Sarradj: Walking noise and its characterization. Proceedings of ICA 2001, 2001. 2 pages.

[14] EPLF NORM 021029–3, Laminate floor coverings – Determination of drum sound generated by means of a tapping machine. EPLF, Association of European Producers of Laminate Flooring, Bielefeld, Germany, 2004.

[15] A.-C. Johansson, P. Hammer, E. Nilsson: Prediction of subjective response from objective measurements applied to walking sound. Acta Acustica united with Acustica **90(1)** (2004) 161–170.

[16] A.-C. Johansson, E. Nilsson, P. Hammer: Evaluation of drum sound with ISO tapping machine. J. Building Acoustics **12(2)** (2005) in press.

[17] B. Watters: Impact-noise characteristics of female hard-heeled foot traffic. Journal of the Acoustical Society of America **37(4)** (1965) 619–630.

[18] W. Scholl, W. Maysenhölder: Impact sound insulation of timber floors: Interaction between source, floor coverings and load bearing floor. J. Building Acoustics **6(1)** (1999) 43–61.

[19] W. Scholl: Impact sound insulation: The standard tapping machine shall learn to walk! J. Building Acoustics **8(4)** (2001) 245–256.

[20] A.-C. Johansson, P. Hammer, J. Brunskog: Tapping machine and footstep sound. Proceedings of 5th Euronoise 2003, Naples, 2003. 5 pages.

[21] A.-C. Johansson, E. Nilsson: Measurement of drum sounds 1636-03 (04207). Technical Report NT573, Nordic Innovation Centre, Oslo, 2005.

[22] J. Brunskog, P. Hammer: The interaction between the ISO tapping machine and lightweight floors. Acta Acustica united with Acustica **89(2)** (2003) 296–308.

[23] E. Zwicker, H. Fastl: Psychoacoustics, facts and models, 2nd edition. Springer-Verlag, Berlin, 1999.

[24] B. C. Moore: An introduction to the psychology of hearing, 5th Ed. Academic Press, London, 2003.

[25] ISO 226 Acoustics – Normal equal-loudness-level contours. 2003.

[26] D. Robinson, R. Dadson: A re-determination for the equal loudness level contours for pure tones. British Journal of Applied Physics **7** (1956) 166–181.

[27] Y. Suzuki, H. Takeshima: Equal-loudness-level contours for pure tones. Journal of the Acoustical Society of America **116(2)** (2004) 918–933.

[28] H. Fastl: The psychoacoustics of sound-quality evaluation. Acta Acustica united with Acustica **83** (1997) 754–764.

[29] S. Stevens: Psychophysics. Introduction to its perceptual, neural, and social prospects. Wiley, New York, 1975.

[30] L. Cremer, H. Müller, T. Schultz: Principles and applications of room acoustics, Volume 1. Applied scinece publishers, Barking, Essex, England, 1982.

[31] H. Fletcher, W. Munson: Relation between loudness and masking. Journal of the Acoustical Society of America **9** (1937) 1–10.

[32] E. Zwicker, B. Scharf: A model of loudness summation. Psychological Review **72** (1965) 3–26.

[33] B. C. Moore, B. Glasberg: A revision of Zwicker's loudness model. Acta Acustica united with Acustica **82** (1996) 335–345.

[34] E. Zwicker: What is a meaningful value for quantifying noise reduction? InterNoise 85, Munich, 1985. 47–56.

[35] E. Zwicker: Meaningful noise measurement and effective noise reduction. Noise Control Engineering Journal **29** (1987) 66–76.

[36] ISO 532: Method for calculating loudness level. 1975.

[37] S. Stevens: The measurement of loudness. Journal of the Acoustical Society of America **27** (1955) 815–829.

[38] S. Stevens: Calculation of loudness of complex noise. Journal of the Acoustical Society of America **28** (1956) 807–832.

[39] S. Stevens: Procedure for calculating loudness: Mark VI. Journal of the Acoustical Society of America **33** (1961) 1577–1585.

[40] E. Zwicker: Subdivision of the audible frequency range into critical bands (frequenzgruppen). Journal of the Acoustical Society of America **33** (1961) 248.

[41] DIN 45631: Procedure for calculating loundness level and loudness. 1991.

[42] E. Zwicker: Procedure for calculationg loudness of temporally variable sounds. Journal of the Acoustical Society of America **62** (1977) 675–682.

[43] R. Hellman, E. Zwicker: Why can a decrease in db(a) produce an increase in loudness? Journal of the Acoustical Society of America **82(5)** (1987) 1700–1705.

[44] B. Berglund: Loudness scaling in environmental psychoacoustics. – In: Subjective and Objective Evaluation of sound. E. Ozimek (ed.). World Scientific Publishing, Singapore, 1990, 3–14.

[45] H. Fastl: Loudness and annoyance of sounds: Subjective evaluation and data from ISO 532B. InterNoise 85, Münich, 1985. 1403–1406.

[46] ASAZ24.3: American tentative standards for sound level meters for measurement of noise and other sounds. Journal of the Acoustical Society of America **8** (1936) 147–152.

[47] E. Zwicker, W. Daxer: A portable loudness meter based on psychoacoustical models. InterNoise 81, Amsterdam, 1981. 869–872.

[48] Y. Ando: A theory of primary sensations and spatial sensations measuring environmental noise. Journal of Sound and Vibration **241(1)** (2001) 3–18.

[49] C. L. Morfey: Dictionary of acoustics. Academic Press, London, 2001.

[50] G. von Bismarck: Sharpness of steady sounds. ACUSTICA **30** (1974) 159–172.

[51] W. Aures: Berechnungsverfahren für den sensorischen Wohlklang beliebiger Schallsignale. ACUSTICA **59** (1985) 130–141.

[52] J. Poggenburg, K. Genuit: How we may succeed in product sound quality. Proceedings of International Congress on Sound and Vibration 1999, Copenhagen, Denmark, 1999. 3055–3062.

[53] R. Guski: Psychological methods for evaluating sound quality and assessing acoustic information. Acta Acustica united with Acustica **83** (1997) 765–774.

[54] M. Bodden: Perceptual sound quality evaluation. Proceedings of InterNoise 2000, Nice, France, 2000. 1699–1704.

[55] NT ACOU 111, Human sound perception — Guidelines for listening tests. Nordtest Method, www.nordtest.org, 2002.

[56] ISO/TS 15666 Acoustics — Assessment of noise annoyance by means of social and socio-acoustic surveys. 2003.

[57] IEC Publication 60268–13, Listening tests on loudspeakers.

[58] D. Hammershøi, H. Møller: Sound transmission to and within the human ear canal. Journal of the Acoustical Society of America **100 (1)** (1996) 408–427.

[59] H. Møller, D. Hammershøi, C. Jensen, M. Sorensen: Evaluation of artificial heads in listening tests. J. Audio Eng. Soc. **47(3)** (1999) 83–100.

[60] A. Järvinen, P. Maijala: On the use of real head recordings in product sound design. Proceedings of InterNoise 97, 1997.

[61] H. Møller: Fundamentals of binaural technology. Applied Acoustics **36(3)** (1992) 171–218.

[62] E. Poulton: Bias in quantifying judgements. Erlbaum, Hove and London, 1989.

[63] A. L. Edwards: Techniques of attitude scale construction. New York, Appleton-Century-Crofts, Inc., 1957.

[64] D. Fucci, D. Harris, L. Petrosino, M. Banks: The effect of preference for rock music on magnitude estimation scaling behavior in young adults. Perceptual and Motor Skills **76** (1993) 1171–1176.

[65] ISO 11056:1999 Sensory analysis – Methodology – Magnitude estimation method. 1999.

[66] M. Khan, O. Johansson, U. Sundbäck: Evaluation of annoyance reponse to engine sounds using different rating methods. Proceedings of InterNoise 96, Liverpool, UK, 1996. 2517–2520.

[67] H. David: The method of paired comparisons. Griffin, London, 1988.

[68] O. Dykstra: Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs. Biometrics **16** (1960) 176–188.

[69] W. Clatworthy: Tables of two-associate-class partially balanced designs. Applied Mathematics Series/National Bureau of Standards, Washington, D.C., 1973.

[70] L. Thurstone: Psychophysical analysis. The American journal of psychology **38** (1927) 368–389.

[71] F. Mosteller: Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. Psychometrika **16** (1951) 3–9, 203–206, 207–218.

[72] R. Bradley: Paired comparisons: Some basic procedures and examples. – In: Handbook of Statistics Vol. 4. Elsevier Science Publisher, North-Holland Publishing Co, Amsterdam, 1984, 299–326.

[73] W. Glenn, H. David: Ties in paired-comparison experiments using a modified Thurstone-Mosteller model. Biometrics **16** (1960) 86–109.

[74] P. Rao, L. Kupper: Ties in paired-comparison experiments: a generalization of the Bradley-Terry model. Journal of the American Statistical Association **62** (1967) 194–204 Corrigenda 63, p.1550.

[75] R. Davidson: On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. Journal of the American Statistical Association **65** (1970) 317–328.

[76] M. Glickman: Parameter estimation in large dynamic paired comparison experiments. Applied Statistics **48** (1999) 377–394.

[77] H. Scheffé: An analysis of variance for paired comparisons. Journal of the American Statistical Association **47** (1952) 381–400.

[78] A. Gabrielsson, H. Sjögren: Perceived sound quality of sound-reproducing systems. Journal of the Acoustical Society of America **65(4)** (1979) 1019–1033.

[79] A. Gabrielsson, B. Hagerman, T. Bech-Kristensem, G. Lundberg: Perceived sound quality of reproductions with different frequency responses and sound levels. Journal of the Acoustical Society of America **88(3)** (1990) 1359–1366.

[80] D. V. Anders Sköld, M. Kleiner: Perceived sound character and objective properties of powertrain noise in car compartments. Acta Acustica united with Acustica **91** (2005) 349–355.

[81] C. E. Osgood: The nature and measurement of meaning. Psychological Bulletin **49** (1952) 197–237.

[82] C. E. Osgood, G. J. Suci, P. H.Tannenbaum: The measurement of meaning. University of Illinois Press, Urbana, 1957.

[83] L. Solomon: Semantic approach to the perception of complex sounds. Journal of the Acoustical Society of America **30** (421–425) 1958.

[84] R. Bisping: Car interior sound quality: Experimental analysis by synthesis. Acta Acustica united with Acustica **83** (1997) 813–818.

[85] I. Sobhi, P. Ladegaard: Design of combination metrics for two household appliances. Proceedings of International Congress on Sound and Vibration 1999, Copenhagen, Denmark, 1999. 3071–3078.

[86] A.Zeiter, J. Hellbruck: Semantic attributes of environmental sounds and their correlation with psychoacoustic magnitudes. 17th International Congress on Acoustics, Rome, 2001.

[87] J. Stevens: Applied multivariate statistics for the social sciences. Lawrence Erlbaum Associates, Inc., Mahwah, N.J., 2002.

[88] StatSoft, Inc. Electronic Statistics Textbook. Tulsa, OK: StatSoft. http://www.statsoft.com/textbook/stathome.html, 2004.

[89] ISO 717: Rating of sound insulation in buildings and of building elements. 1996.

[90] K. Bodlund: Alternative reference curves for evaluation of the impact sound insulation between dwellings. Journal of Sound and Vibration **102(3)** (1985) 384–402.

[91] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold: Multi- and megavariate data analysis: Principles and applications. Umetrics, Umeå, 2001.

[92] N. H. Timm: Applied multivariate analysis. Springer-Verlag, New York, 2002.

[93] D. C. Montgomery: Design and analysis of experiments, 5th Ed. John Wiley & Sons, New York, 2001.

[94] MATLAB. Statistics toolbox, help, Release 13, 2002.

[95] C. Guastavino, B. Katz: Perceptual evaluation of multi-dimensional spatial audio reproduction. Journal of the Acoustical Society of America **116** (2004) 1105–1115.

[96] N. Zacharov, K. Koivuniemi: Audio descriptive analysis and mapping of spatial sound displays. Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, 2001. 95–104.

[97] O. Johansson: Experimental and multivariate analysis methods for sound quality evaluation of diesel engines, 1996/203D. Dissertation. Luleå University of Technology, Luleå, 1996.

[98] M. S. Khan: Sound quality evaluation of heavy-duty engines in free field conditions, DT98/16. Dissertation. Luleå University of Technology, Luleå, 1998.

[99] T. F. Cox, M. A. Cox: Multidimensional Scaling, 2nd Edition. Chapman & Hall, Boca Raton, 2001.

[100] W. S. Torgerson: Multidimensional scaling: I Theory and method. Psychometrika **17** (1952) 401–419.

[101] P. Susini, S. McAdams, S. Winsberg: A multidimensional technique for sound quality asessment. Acta Acustica united with Acustica **85** (1999) 650–656.

[102] D. V. Anders Sköld, M. Kleiner: Predicting consumer preference from expert sensory ratings of sounds. Proceedings of ICA 2004, Kyoto, Japan, 2004. 1825–1828.

Included papers

A

Aspects on paired comparison models for listening tests

Ann-Charlotte Johansson, Per Hammer, Erling Nilsson
Engineering Acoustics, LTH, Lund University, Box 118, 221 00 Lund, Sweden. Email: Ann-Charlotte.Johansson@acoustics.lth.se

Summary
In many acoustic environments, for example, buildings or vehicles, as well as in product development, etc., there is a need to rank and classify sounds. A frequently used procedure is the paired comparison test. A number of ways to perform and analyse this test exist. In this paper a comparison of different existing approaches is made. The main focus is set on the basic models by Thurstone-Mosteller and Bradley-Terry. Extensions to both of the models, concerning ties, are presented along with a discussion of when they should be used. Thereafter, procedures to test whether the calculated ranking values are statistically different are presented. The advantages and disadvantages of these methods are discussed, and some examples are given which consider the responses from tests on drum sound from floor coverings. It is seen that the choice between these models is not crucial. Ties are generally recommended as they add information and can decrease the results' confidence intervals/regions. The model by Bradley-Terry and its extensions are recommended. However, if only scale values are requested, the treatments are somewhat similar in character and no ties are allowed, the Thurstone-Mosteller model is recommended due to the simplicity of the calculations.

PACS no. 43.66

1. Introduction

In many acoustic environments, for example, buildings or vehicles, as well as in product development etc., there is a need to rank and classify sounds. In [1] a review of some psychological methods for evaluating sound quality (mainly response scales) is presented. Scales have the advantage in that they can produce an absolute value corresponding to a certain sensation. However, scales can be hard for an assessor (the listener) to use — uncertainties about whether the assessors have used and understood the scale equally can occur — and it can therefore be difficult for the researcher to analyse the result. A way of avoiding this problem is to use paired comparison tests. In such a test the assessors are asked to tell which of two treatments — or stimuli or sounds — has a certain attribute (such as a pleasant sound). The need for the assessor to have a longer term memory is eliminated and more consistent answers can be achieved. In [2] the method of paired comparison, the method of equal-appearing intervals and the method of successive intervals were compared in the evaluation of annoyance response to engine sounds. It was concluded that consistent judgements of annoyance were observed by the paired comparison data. Judgements of annoyance using the methods of equal-appearing intervals and successive intervals were made consistently only by trained assessors. Recently in [3] various listening-test methods were compared. It is concluded that the discrimination power was greater for the paired comparison test than for the methods using scales.

One disadvantage of paired comparison models is that no absolute values corresponding to a certain attribute are given, but only relative values. Another disadvantage is the rapid growth of comparisons needed when the number of treatments increases. This could, however, be treated to some extent by the use of incomplete balanced design where not all pairs are compared in the test. Scales and paired comparison methods can complement each other and should not be seen as two alternatives; the choice should be based on how the result is to be used.

In [4] a paired comparison model allowing ties, i.e. allowing treatments to be declared equal, by Rao and Kupper [5] is used. Questions regarding the various methods to analyse the result arose as there are a number of ways to perform and analyse paired comparison tests. In [3] the scores from the paired comparison test were reported to be similar when using different methods to analyse the result. Is that a coincidence? The intention of this paper is to increase the knowledge of these models, point out the main ideas behind these models and clarify their differences and applicability in listening tests.

Two major basic pair comparison methods exist today, the Thurstone-Mosteller [6, 7] model and the Bradley-Terry model [8, 9]. These models are both so-called linear models. The linear paired comparison model is defined by David [10] as

$$P(T_i \rightarrow T_j) = H(V_i - V_j);$$

that is, the probability of choosing treatment T_i when compared with T_j is a function of the difference in their strengths (or scores) V_i and V_j only, where the function H is a symmetric cumulative distribution function[1]. The Bradley-Terry model assumes a standard logistic distribution function and the Thurstone-Mosteller model assumes a normal (Gaussian) distribution function.

Sometimes the assessors are not able to discern any difference in the pair. This is the case when the treatments

[1] In statistics a symmetric cumulative distribution function has the properties that $H(-x) = 1 - H(x)$

are equal in the specific attribute assessed or when the difference is too small to be perceived. When ties are not allowed, the assessors are forced to make a selection and the choice is, hopefully, made randomly. When no models for handling ties were developed, a way of excluding this random behaviour from the input data was to allow ties in the test but to ignore them in the analysis. Even though models that do allow ties in the analysis were developed during the 60's and 70's [11, 5, 12], there are situations where they are not used today although they could increase the information in the analysis.

The paired comparison methods result in a magnitude rating of the included treatments regarding the attribute that the question concerns. In the last section of this paper, statistical methods are presented to check whether the observed differences are significant.

An attempt has been made to maintain consistency in notation within this paper rather than to follow the notations of the original papers.

The models will be presented with the result of three test examples described in the following section to show the similarities and differences between the models and to enable verifying calculations. In section 3 models excluding ties are presented, starting with the Thurstone-Mosteller model [6, 7], followed by the Bradley-Terry model [8, 9] and the Gamma paired comparison models by Stern [13, 14] who shows that the models by Thurstone-Mosteller and Bradley-Terry can be described in a general way using gamma random variables. In section 4 models including ties are presented including an extension of the Thurstone-Mosteller model by Glenn and David [11], and two extensions of the Bradley-Terry model by Rao and Kupper [5] and Davidson [12]. Methods to estimate whether or not there are significant differences in the stimuli's scores are presented in section 5. An analytical formulation to find the level of significant difference in a pair is presented for the models based on the Bradley-Terry model. In the final section conclusions and recommendations to design a paired comparison test and to select an evaluation model are given.

It is seen that the choice of either Thurstone-Mosteller or Bradley-Terry is not crucial. As the former model provides an algebraic solution, it is recommended when scale values are requested. However, when estimations of their differences are to be made, when the treatments are inhomogeneous or if the design is either incomplete or unbalanced, the latter model is recommended. When preference is the objective, ties should be allowed as they add information. The confidence intervals of the treatment ratings, calculated from the data in the test examples, are shorter when ties are allowed. The use of ties enabled the discovery that the sampling distribution of the normal for one pair of the sounds is binormal; a test without ties would conceal that information. The choice between the Rao-Kupper and Davidson models in terms of their ability to handle ties is not critical.

Table I. Example 1: The frequency, a_{ij}, with which the row i drum sound, DS, is assessed louder than column j drum sound; no ties were allowed. 24 assessors participated.

DS	A	B	C	D	E
A	-	13	22	22	22
B	11	-	24	24	21
C	2	0	-	18	13
D	2	0	6	-	10
E	2	3	11	14	-

Table II. Example 2: Listening test allowing ties with 24 participant assessors. In the top matrix the frequencies, a_{ij}, with which the row i drum sound, DS, is assessed louder than column j drum sound are shown; the lower matrix shows the reported number of ties, a_{ij0}.

DS	A	B	C	D	E
A	-	14	21	21	21
B	9	-	14	22	17
C	3	0	-	8	7
D	2	0	4	-	2
E	3	1	2	4	-
A	-	1	0	1	0
B	-	-	10	2	6
C	-	-	-	12	15
D	-	-	-	-	18
E	-	-	-	-	-

2. Test examples

This section data from three listening tests on drum sound are presented. The data will be used to show the similarities and differences between the various paired comparison models and to enable verifying calculations.

In a listening test reported in [15] 24 assessors were comparing the loudness of five recorded drum sounds. The question was: "Which of the sounds is louder (Swedish *ljudstarkast*)?" Drum sound refers to the sound that occurs when an object, e.g. a foot, strikes the flooring in the same room as the receiving ear. The drum sounds were recorded with a person walking on various laminate floor coverings. The presentation order of the comparisons was random. The assessors came twice within three days with at least 4 hours in between each time. One half started by comparing the sounds without ties permitted, hence they were forced to make a selection; the second time they participated ties were allowed. The other half performed the test in the reversed order. In table I the result from the listening test when no ties were allowed is shown. In table II the test results when ties were permitted are shown. The results are given in frequencies a_{ij} with which the row drum sound i is assessed louder than column drum sound j; a_{0ij} denotes the number of ties. Proportion of times that the row drum sound i is assessed louder than column drum sound j, p_{ij}, are then given by $p_{ij} = a_{ij}/n_{ij}$ where n_{ij} is the total number of comparisons made for the pair (i, j), $n_{ij} = a_{ij} + a_{ji}(+a_{0ij})$.

Table III. Example 3: Listening test allowing ties with 31 participant assessors. In the top matrix the frequencies, a_{ij}, with which the row i drum sound, DS, is assessed louder than column j drum sound are shown; the lower matrix shows the reported number of ties, a_{ij0}.

DS	A_3	B_3	C_3	D_3	E_3	F_3	G_3
A_3	-	8	15	29	7	31	1
B_3	4	-	15	30	3	29	3
C_3	0	0	-	25	0	29	0
D_3	1	0	1	-	1	21	0
E_3	9	13	23	30	-	31	0
F_3	0	1	0	1	0	-	1
G_3	25	24	29	31	24	30	-
A_3	-	19	16	1	15	0	5
B_3	-	-	16	1	15	1	4
C_3	-	-	-	5	8	2	2
D_3	-	-	-	-	0	9	0
E_3	-	-	-	-	-	0	7
F_3	-	-	-	-	-	-	0
G_3	-	-	-	-	-	-	-

In another listening test, reported in [16], 31 assessors compared the loudness of seven recorded drum sounds, not the same as above. The same question was used: "Which of the sounds is louder?" The presentation order of the comparisons was random. Ties were permitted. In table III the test results are shown.

3. Models excluding ties

The method of paired comparisons is found as early as 1860 when Fechner published Elemente der Psychophysik [17], translated to English in 1965 [18]. Fletcher defined psychophysics as "an exact science of the functional relations of dependency between body and mind". Fletcher included in psychophysics the measurement and quantification of the perception, in order to find the correlation between the psychological scales and the physical measurements of the stimuli [19]. Fletcher used vessels and assessed which, out of two, was the heaviest. The variability of the apparent mass was determined by assuming the mass to be normally distributed about the true mass [10]. Thurstone [6, 20] used these psychophysical scaling methods for measurements of psychological stimuli that could not be measured by physical methods, and published in 1927 a paper which, with extensions made by Mosteller [7], formed one of the methods for paired comparisons used today.

3.1. Thurstone-Mosteller model

Thurstone's *law of comparative judgement* assumes that every given stimulus, T_i, is associated with a sensation, X_i, and that this process can be ordered on a psychological scale. How a group of stimuli is ordered depends on the attribute that is of interest. However, even though the attribute is set, a given stimulus does not always produce the same sensation but fluctuates around its "true" scale value, S_i = mean of X_i. The scale is then defined so that the fluctuation forms a normal distribution. In figure 1 the distributions of the produced sensation for two stimuli on a psychological scale are shown. It is seen that when stimuli T_i and T_j are compared T_i will usually be given a higher degree than T_j of the attribute that is of interest, but the reverse, $X_j > X_i$, might happen. Actually, this has to happen in order to have anything but a rank order.

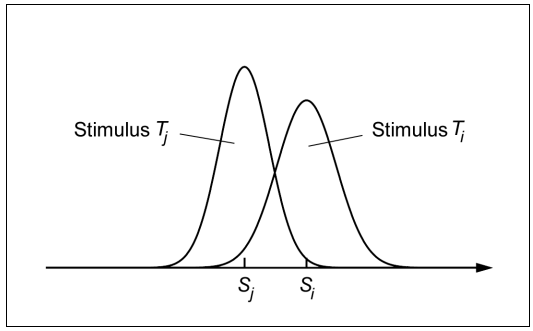

Figure 1. The distributions of the produced sensations for two stimuli on a psychological scale.

The difference between T_i's and T_j's scale value is given by

$$S_i - S_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j}, \tag{1}$$

where S_i, S_j and σ_i, σ_j are the scale values and standard deviations for stimuli T_i and T_j respectively. z_{ij} is the normal deviate corresponding to the proportion of times T_i is selected over T_j, i.e. $z_{ij} = \Phi^{-1}(p_{ij})$ when Φ is the normal cumulative distribution function and $p_{ij} = a_{ij}/n_{ij}$. r_{ij} is the correlation coefficient between T_i and T_j. In figure 2 the distribution of the sensation difference, $X_i - X_j$, is shown. The most common difference (the mean value) is assumed to be the true difference. The shaded area represents the proportions of assessments T_i stronger than T_j, p_{ij}, and it is that information that is received in a comparison test. It is thus from these proportions that estimates of the scale values are to be set and the fitting is made so that the scale values best satisfy the observed proportions. Estimates are here denoted with an "*"; S_i^* is for example the estimate of S_i.

Based on equation (1) Thurstone lists five cases with increasing degrees of assumption and simplification. In all of these cases comparisons $T_i - T_i$ are not allowed; that is, no comparison of the treatment with itself is allowed.

Case I

If the correlation was allowed to vary for each comparison the problem would be unsolvable; therefore the correlation coefficient, r, is in Case I set to be constant throughout the test for the single assessor, i.e. $r_{ij} = r$. At least five stimuli are, however, needed to make the system solvable.

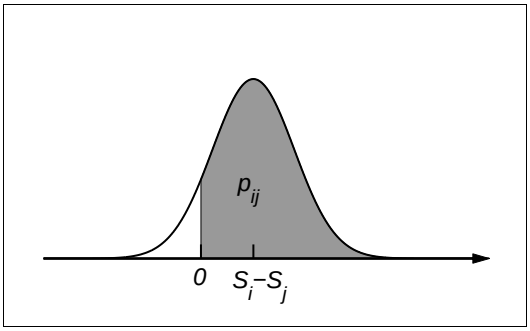

Figure 2. Distribution of the sensation difference, $X_i - X_j$.

Case II
In Case II the conclusions are drawn based on the result from a group of assessors. The assumption is then made that the distribution of the degree of an attribute perceived by a group of assessors is normal. The assumption of Case I applies in II as well.

Case III
When the stimuli are homogenous so that the distracting attributes are few, it can be assumed that the correlation coefficient between the stimuli is low. In Case III it is set to zero; that is, it is assumed that the decision made about one stimulus does not affect the decision about the other stimuli. The assumptions of Case II remain in Case III.

$$S_i - S_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2} \tag{2}$$

At least five stimuli are needed to make this system solvable.

Case IV
In Case IV it is assumed that the standard deviations are similar so that they can be related as

$$\sigma_j = \sigma_i + d$$

where d is assumed to be smaller than σ_i and preferably less than $0.5\sigma_i$. Equation (1) can then be approximated to

$$S_i - S_j = 0.707 z_{ij}(\sigma_i + \sigma_j).$$

At least four stimuli are needed to make this system solvable. As both σ_i and σ_j are still needed, the work needed in solving the equations is not very much reduced, so it is recommended to use Case III instead of Case IV.

Case V
The most commonly used case is Case V. It is here assumed that all standard deviations are equal. Thurstone assumed in [20] that the correlation coefficient r is zero, but Mosteller later showed that this restriction can be limited to be constant without any changes in solution method [7]. The model is now:

Table IV. Design of the proportion matrix showing the proportions of assessments T_i (row) stronger than T_j (column)

	T_1	T_2	...	T_t
T_1	–	$p_{1>2} = p_{12}$	...	$p_{1>t} = p_{1t}$
T_2	$p_{2>1} = p_{21}$	–	...	p_{2t}
...	...	...	–	...
T_t	$p_{t>1} = p_{t1}$	p_{t2}	...	–

$$S_i - S_j = z_{ij}\sqrt{2\sigma^2(1-r)}$$

As $2\sigma^2(1-r)$ can be seen as a scale factor that, without loss of generality, can be set to unity, we have

$$S_i - S_j = z_{ij} \tag{3}$$

Mosteller also showed that Case V can be written

$$P(T_i \rightarrow T_j) = P(X_i > X_j) = \frac{1}{\sqrt{2\pi}} \int_{-(S_i - S_j)}^{\infty} \mathrm{e}^{-\frac{1}{2}y^2} \mathrm{d}y \tag{4}$$

and that the approach Thurstone presents in e.g. [21] is a least square solution of the stimulus positions on the sensation scale. In the next section the least square solution will be presented; however, maximum likelihood estimates of the scale values can also be used as discussed in the following section.

3.1.1. Calculation procedure — Least square solution

From the paired comparison test the p_{ij} values as in table IV are given. A way to proceed with the calculation is to first apply Case V and then test the goodness of fit with a chi-square test. In the goodness to fit test the model's ability to describe the original data from the listening test is tested. If there is reason to believe that the Case V model is inappropriate, the Case III model should be tested and checked with another chi-square test. (Case IV is not worth trying in between as the work input is the same as for Case III.) How to proceed with the more complex Cases I and II, if needed, is not described here. However, an approximation of Case I can be found in [22].

Case V, complete data
After arrangement of the proportion matrix the normal deviates of each p_{ij} are calculated. Estimates of the scale values, S_i^*, $i = 1, ..., t$, for the t stimuli assuming Case V are then received as

$$\begin{aligned} \sum_{j \neq i}^{t} z_{ij} &= (t-1)S_i - \sum_{j \neq i}^{t} S_j \\ &= (t-1)S_i - [0 - S_i] = tS_i \\ S_i^* &= \frac{1}{t}\sum_{j \neq i}^{t} z_{ij}. \end{aligned} \tag{5}$$

Table V. Matrix of successive row differences of the normal deviates of the values in table IV.

	1	2	...	t
$2-1$	$z_{21}-z_{11}$	$z_{22}-z_{12}$	...	$z_{2t}-z_{1t}$
$3-2$	$z_{31}-z_{21}$	$z_{32}-z_{22}$	...	...
...	...	...	...	...
$t-(t-1)$	$z_{t1}-z_{(t-1)1}$	...	...	$z_{tt}-z_{(t-1)t}$

where the sum of the scale values is set equal to zero, i.e. the average scale value is equal to zero.[2] A negative scale value then indicates that the stimulus is assessed as less strong than the average and a positive thus indicates that the stimulus is stronger than the average. If wanted, a positive scale can be achieved by adding a constant to the scale values as the origin is arbitrary.

In the solution above it is assumed that we have a complete set of data. When some information is missing another approach is needed; see the section below.

A drawback of the law of comparative judgement is that when the proportions of an assessment, p_{ij}, tend to zero or one, large numbers of z_{ij} are then introduced and in the limit z_{ij} becomes indeterminate. Mosteller recommends that these entries in the matrix be excluded whenever $|z_{ij}|$ exceeds 2, which corresponds to a $p_{ij} > 0.977$ or $p_{ij} < 0.023$ [7]. Edwards suggests that when the number of assessors is large, more than 200, p_{ij} values of 0.99 or 0.01 might be used; otherwise his recommendation is in line with Mosteller's [23]. The same approach as with incomplete data should thereafter be taken.

Case V, incomplete data

The approach presented follows Edwards [23].

1. Arrange the stimuli in approximately ranking order. An approximate order can be arranged by calculating the row or column sum.
2. Calculate the normal deviates z_{ij} (exclude entries whenever $|z_{ij}|$ exceeds 2.
3. Calculate the successive differences of the row entries as presented in table V. The successive differences are only calculated whenever both entries are known. Consider diagonal entries z_{ii} as known (value = 0).
4. Sum each row and divide this sum with the number of contributing equations, k, for that row. $\sum_{j=1}^{t}(z_{ij}-z_{(i-1)j})/k = D_{i(i-1)}$
5. Estimates of the scale values are given by cumulatively adding the mean values as $S_i^* = S_{i-1}^* + D_{i(i-1)}, S_1^* = 0, i = 2 \ldots t$

Case V, goodness to fit

To check that the assumptions of Case V are right, a goodness to fit test can be made. A chi-square test presented by Mosteller [7] is given here.[3] The null hypothesis, H_0, is to test if the model of Case V is true and the alternative hypothesis, H_a, is the model is not true for some $i, j, i \neq j$.

$$H_0 : p_{ij}^* = p_{ij} \quad i \neq j, \quad i,j = 1,...,t$$
$$H_a : p_{ij}^* \neq p_{ij} \quad \text{for some} \quad i,j,i \neq j$$

Mosteller then uses the inverse sine transformation to take into account that even for rather large n_{ij} sampling distribution proportions close to 1 are not normal. The sampling distrubution of the statistic $\arcsin\sqrt{p_{ij}}$ is, however, normal. The hypothesis is tested on a chosen significance level, α. α is the probability to reject a hypothesis even though it is true and making a so-called Type I error [25]. By lowering α we are setting a stronger demand on our hypothesis. The goodness to fit is calculated by

$$\chi^2 = \sum_{i<j} \frac{\left(\arcsin\sqrt{p_{ij}^*} - \arcsin\sqrt{p_{ij}}\right)^2}{821/n_{ij}} \tag{6}$$

where p_{ij}^* is the estimated proportions and p_{ij} is the observed proportions, $p_{ij} = a_{ij}/n_{ij}$. p_{ij}^* is given by $p_{ij}^* = \Phi(z_{ij})$ where Φ is the normal cumulative distribution function and z_{ij} is calculated using the estimated scale values in equation (3). $\arcsin\sqrt{p_{ij}}$ is measured in degrees; the value $821/n_{ij}$ is approximately the standard deviation of $\arcsin\sqrt{p_{ij}}$. n_{ij} is the number of comparisons per stimulus pair (i.e. equal to the number of assessors if each assessor compares every pair once).

The calculated chi square is then used to check the significance of the discrepancies. The number of degrees of freedom for a test with complete data is $(t-1)(t-2)/2$. (There are $t(t-1)/2$ pairs and $(t-1)$ estimates.) For an incomplete data set, the degrees of freedom are reduced further by the number of absent comparisons. It is recommended to always report the P-value of the significance, that is, the smallest level of significance that would lead to rejection of the null hypothesis. A high P-value then indicates high likelihood of doing wrong if the null hypothesis is rejected. Guilford [24] recommends using the 0.01 level to reject the hypothesis that the model is tenable. If the P-value is considered too low, e.g. < 0.01, there is something wrong with the assumptions of normality, unidimensionality or equal standard deviation. An effect of nonunidimensionality is circular triads, that is, A is chosen over B, B is chosen over C but C is chosen over A. If circular triads are suspected, methods to determine the number of circular triads can be found in [10, 26, 27]. In an investigation concerning circular errors using sounds recorded in a duplex high speed train (TGV) it is concluded that the result using assessors with a rate of mistakes of less than 10% (11 persons) is only slightly affected compared to when using all assessors (35 persons) regardless of their rate of mistakes; the rate of mistakes should instead be used as an indication of

[2] The row subscript is here given first and the column second. Please note that in the references concerning the Thurstone-Mosteller model, the matrix notation is reversed so that the summation is made for each column. The upper limit is in this paper always t and is from now on omitted in the equations.

[3] Although Guilford [24] uses Mosteller's chi-square test he points out that such a test assumes zero correlation between the stimuli.

the difficulty of the test [26]. If the problem of poor goodness to fit is unequal standard deviation, it is solved by applying Case III. It has, however, been commented by several authors that this test is unsensitive to the normality assumptions [7, 24] and accepts too easily the model in general [10]. Mosteller has shown that if the standard deviation of one stimulus is different it will only affect the scale value of that stimulus, and if it is scaled close to the mean of the scale with the Case V model, the influence of the different standard deviation is slight [7].

Case III

Burros presented the following approach to receive estimates of the standard deviation [28]. In the development it is assumed that the standard deviations are still somewhat similar.

The variance, V_i, of the normal deviates of each row i in table IV is given by

$$V_i = \sqrt{\frac{\sum_{j=1}^{t} (z_{ij} - \bar{z}_i)^2}{t}}$$

where $\bar{z}_i$ is the mean of the z_{ij} values in row i. The standard deviations can then be estimated by

$$\sigma_i^* = \frac{2t\,(1/V_i)}{\sum_{i=1}^{t} 1/V_i} - 1. \tag{7}$$

Every normal deviate of the proportions in table IV should then be multiplied with $\sqrt{\sigma_i^{*2} + \sigma_j^{*2}}$. The procedure described for Case V (complete or incomplete data) can thereafter be applied.

Case III, goodness to fit

By the use of equation (2) and the estimates $S_i^*, i = 1, ..., t$, estimates p_{ij}^* of p_{ij} can be retrieved from $p_{ij}^* = \Phi(z_{ij})$. A similar procedure to that in Case V is thereafter applied, but the number of degrees of freedom for a complete set of data is reduced to $(t-1)(t-4)/2$ as another $(t-1)$ degrees of freedom is lost in the estimation of the standard deviations. For an incomplete set, the degrees of freedom is reduced further by the number of absent comparisons.

As the degree of freedom is decreased it has been noticed by Guilford that the goodness to fit of Case III might decrease compared with Case V [24]. The test is thus sometimes undersensitive and sometimes oversensitive, but it is nevertheless recommended to perform this test.

3.1.2. Calculation procedure — Maximum likelihood solution

In the preceding section a least square solution was presented. It is the most popular approach as the solution is algebraic. A more statistically efficient approach is, however, given the use of maximum likelihood estimates. The number of times T_i is selected over T_j is then assumed to be a sample of a binomial distribution and estimates of the scale values are found by maximizing the likelihood function, L_T

$$L_T = \prod_{i<j} \binom{n_{ij}}{a_{ij}} p_{ij}^{a_{ij}} \left(1 - p_{ij}\right)^{n_{ij} - a_{ij}} \tag{8}$$

where a_{ij} is the number of times T_i was selected over T_j and $n_{ij} = a_{ij} + a_{ji}$. p_{ij} is for the Case V model the normal deviate of $(S_i - S_j)$ and for Case III the normal deviate of $(S_i - S_j)/\sqrt{\sigma_i^2 + \sigma_j^2}$.

In [29] a Monte Carlo EM algorithm to find the maximum likelihood estimation, evolved for the case when the number of items is large, is demonstrated.

In this paper, however, only the least square, algebraic, solution is used.

3.2. Bradley-Terry model

In 1929 Zermelo presented a paper in which he wonders how to evaluate chess players in a tournament where no player had yet met any other [30]. If the strength of every player is denoted π_i, the probability of a win when player i meets j would, according to Zermelo, be

$$\pi_{ij} = \frac{\pi_i}{\pi_i + \pi_j}.$$

Zermelo then presents the 'combined probability' (i.e. the likelihood function) and maximizes it to obtain estimates of the π_i. He comments that the ranking of the players is the same as when the score is given by the number of wins as long as the tournament is balanced, $n_{ij} = n$, i.e. the number of comparisons per stimulus pair is constant. Zermelo's work was not used in applications other than tournaments, however, and it was not until Bradley and Terry presented their similar but more extensive work that the model became more widely known.

The Bradley-Terry model [8, 9] gives maximum-likelihood estimates of the treatment ratings using a generalization of a binomial model and distribution. The probability of choosing T_i when compared to T_j is given as

$$P(T_i \rightarrow T_j) = \frac{\pi_i}{\pi_i + \pi_j} =$$

$$= \frac{1}{4} \int_{-(\ln \pi_i - \ln \pi_j)}^{\infty} \operatorname{sech}^2(y/2)\mathrm{d}y \tag{9}$$

where $\pi_i, ..., \pi_t$, are the treatment ratings representing relative selection properties for the t treatments, $i \neq j, i, j = 1, ..., t$. As for the Thurstone-Mosteller model, comparisons T_i with T_i are not allowed. The probability can also be described as an integral of the squared hyperbolic secant as shown above, where the probability is seen to be dependent on the natural logarithm of the π-values of the treatments. This is the reason why comparison of the treatments should be made on the natural logarithm of the treatment ratings, by Bradley and Terry called the true merits.

3.2.1. Calculation procedure

The binomial component of the likelihood function is here

$$\binom{n_{ij}}{a_{ij}} \left(\frac{\pi_i}{\pi_i + \pi_j}\right)^{a_{ij}} \left(\frac{\pi_j}{\pi_i + \pi_j}\right)^{a_{ji}}$$

where a_{ij} is the number of times T_i was selected and a_{ji} is the number of times T_j was selected ($a_{ij} + a_{ji} = n_{ij}$). The complete likelihood function, L_B, then becomes

$$L_B = \prod_{i<j} \binom{n_{ij}}{a_{ij}} \frac{\prod_{i=1}^{t} \pi_i^{a_i}}{\prod_{i<j} (\pi_i + \pi_j)^{n_{ij}}} \tag{10}$$

where $a_i = \sum_{j,j\neq i} a_{ij}$. By maximizing the natural logarithm of L_B and using the constraint $\sum_i^t \pi_i = 1$, estimates π_i^* of π_i are obtained:

$$\frac{a_i}{\pi_i^*} - \sum_{j,j\neq i} \frac{n_{ij}}{\pi_i^* + \pi_j^*} = 0, \quad i = 1, ..., t$$

and

$$\sum_i \pi_i^* = 1$$

The solution is received iteratively. Starting values may be $\pi_i^{*(0)} = 1/t$. Estimate π_i^* of π_i is then

$$\pi_i^{*(k)} = \hat{\pi}_i^{(k)} \Big/ \sum_i \hat{\pi}_i^{(k)}$$

where

$$\hat{\pi}_i^{(k)} = a_i \Big/ \sum_{j,j\neq i} \left[n_{ij} \Big/ \left(\pi_i^{*(k-1)} + \pi_j^{*(k-1)} \right) \right]$$

True merits of the treatment are thereafter achieved by taking the natural logarithm of the final treatment ratings.

What Zermelo remarked about the agreement in a balanced experiment of the ranking based on the maximum-likelihood estimates to the ranking based on the total number of selections of one treatment, a_i, Ford has shown in [31]. However Ford also concludes that when the experiment is not balanced the estimates π_i^* still provide an appropriate ranking, which the a_i do not.

Unlike the Thurstone-Mosteller model, there is no problem in the Bradley-Terry model if, for some pairs, one of the treatments is always chosen, i.e. when $p_{ij} = 1$. Still, as for the Thurstone-Mosteller model, problems arise when one or several of the treatments are always chosen in favour of all the others. Whenever this is the case, the treatments must be separated into smaller groups and analysed separately. However, this is seldom a problem as the treatments often are somewhat similar.

Goodness to fit

Bradley presents in [9] two likelihood-ratio tests to check the adequacy of the model. The null hypothesis is

$$H_0 : \pi_{ij} = \pi_i/(\pi_i + \pi_j) \quad i \neq j, \quad i, j = 1, ..., t$$

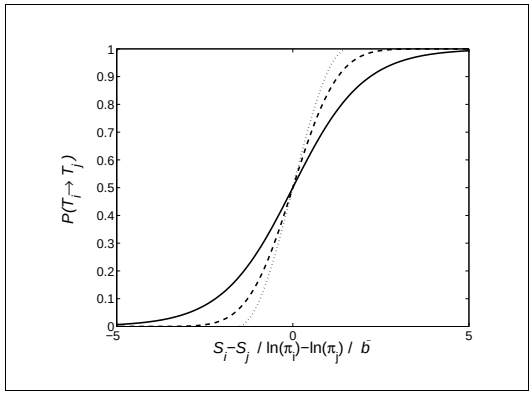

Figure 3. The standard normal (dashed), standard logistic (solid) and angular, section 4.1, (dotted) cumulative distribution function.

and the alternative hypothesis is

$$H_a : \pi_{ij} \neq \pi_i/(\pi_i + \pi_j) \quad \text{for some} \quad i, j, i \neq j.$$

As the methods presented are asymptotically equivalent only one of them is presented here. The statistic is

$$\chi^2 = \sum_{i\neq j} \frac{n_{ij}\{p_{ij} - [\pi_i^*/(\pi_i^* + \pi_j^*)]\}^2}{[\pi_i^*/(\pi_i^* + \pi_j^*)]}$$

where $p_{ij} = a_{ij}/n_{ij}$ and π_i^*, π_j^* are the estimates of $\pi_i, \pi_j, i, j = 1, ..., t$. For large n_{ij}, χ^2 has a chi-square distribution with $\frac{1}{2}(t-1)(t-2)$ degrees of freedom if the null hypothesis is true. Terms with small n_{ij} should be omitted (and the numbers of freedom reduced by one for each pair omitted) according to [32]. Bradley omits terms with $n_{ij} = 0$ in [9].

3.3. Stern — Gamma paired comparisons models

As can be seen by comparing equations (4) and (9), the models by Thurstone-Mosteller and Bradley-Terry have similarities. The Thurstone-Mosteller model assumes a normal (Gaussian) distribution function whereas the Bradley-Terry model assumes a standard logistic distribution function. Their cumulative distribution functions can be seen in figure 3. The probability $P(T_i \rightarrow T_j)$ is seen to be dependent on the difference in merits $(S_i - S_j)$ and $(\ln \pi_i - \ln \pi_j)$ respectively.

Stern has shown that these models can be described in a general way using gamma random variables and that the two models above are special cases of his approach [13, 14]. The outcome of a paired comparison is in his model determined by comparing the waiting time for r events to occur in each of the two processes. The starting point is that a player, or treatment T_i, scores points according to a Poisson process with rate π_i. The scoring process for the different players is assumed to be independent. The time until player T_i scores 1 point is then an exponential

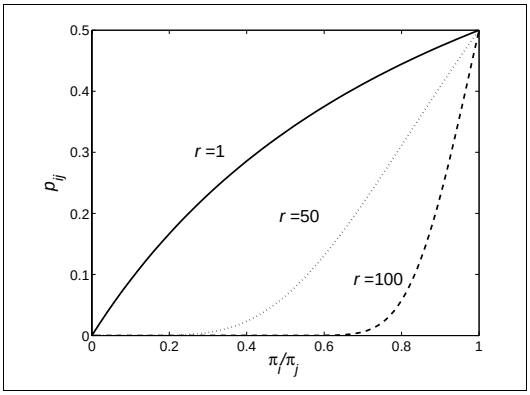

Figure 4. Preference probabilities i gamma paired comparison model as function of the ratio of the scale parameters π_i/π_j.

random variable, or equivalently a gamma random variable with shape parameter $r = 1$, i.e. $X_i \sim \Gamma(1, \pi_i)$. The probability that player T_i scores before T_j is then the probability that $X_i \sim \Gamma(1, \pi_i)$ is less than $X_j \sim \Gamma(1, \pi_j)$ for independent random variables X_i,X_j. This probability is shown to be the Bradley-Terry probability as in equation (9). By comparing gamma random variables with shape parameters other than one, other gamma paired comparison models are obtained. The probability that treatment T_i is chosen over T_j is then

$$P(T_i \rightarrow T_j)^{(r)} = P(X_i < X_j) =$$

$$= \int_0^\infty \int_0^{x_j} \frac{\pi_i^r x_i^{r-1} \mathrm{e}^{-\pi_i x_i}}{\Gamma(r)} \frac{\pi_j^r x_j^{r-1} \mathrm{e}^{-\pi_j x_j}}{\Gamma(r)} \mathrm{dx_i dx_j}. \quad (11)$$

This can be shown to be a function of the shape parameter r and the ratio of the scale parameters π_i/π_j. As mentioned above, $r = 1$ equals the Bradley-Terry model. Stern also shows that as $r \rightarrow \infty$, with $\pi_j = \pi_i + \Delta r^{-1/2} \pi_i$, the Thurstone-Mosteller model is achieved. The behaviour of this probability with different shape parameters is shown in figure 11. It can be seen that for increasing r, a small difference in the scale values gives larger differences in p_{ij}.

To answer the question of whether all models that can be seen as gamma paired comparison models are the same, Stern calculates the goodness to fit using a likelihood ratio test on various sport data sets [14]. Analysing models with r = 0.1, 1, 10 and 50, the fit is seen to be similar for all models and no model is always better than all the others. As the models differ most for extreme values of p_{ij}, Stern generates such data using $r = 0.1$ and then checks how many comparisons are needed using the $r = 50$ model to significantly distinguish between the models. The number was then 250! He concludes that the paired comparison analysis is not very sensitive to the choice of distribution within the class of linear models. It is more important to determine whether a linear model is appropriate.

Table VI. Scale values, S_i^*, true merits, $\ln(\pi_i^*)$, and goodness to fit P-value using the Thurstone-Mosteller model (Cases V and III, algebraic solution to incomplete model) and Bradley-Terry model.

Drum sound	Case V	Case III	B-T
A	1.60	2.70	-0.82
B	1.46	2.39	-0.76
C	0.42	0.90	-3.24
D	0	0	-3.99
E	0.25	0.61	-3.38
P-value	0.79	-	0.35

3.4. Example

With the data from the test example, scale values using the Case V and III models of Thurstone-Mosteller (algebraic solution) and true merits using the Bradley-Terry model are calculated. These results are shown in table VI along with the goodness to fit data. As the data are incomplete, p_{ij} is one and zero at some comparisons, the incomplete model for the Thurstone-Mosteller model is used. The order of the data then becomes important. The order D, E, C, A, B, corresponding to the order of increasing row sum of table I, gave in this case the highest P-value and is shown in table VI. The P-value for Case III could not be calculated as the degrees of freedom was zero; as mentioned before, five stimuli is the least number of stimuli solving the equations for Case III, giving 10 unknowns and 10 equations to solve. The true merits, $\ln(\pi_i^*)$, are calculated until the relative change in the true merits from one iteration to the next is less than 0.001, which resulted in 31 iterations. In figure 5, the scale values of both Case V and Case III are plotted against the true merits. Least-squares fit lines are also added. As the data fit well to the least-squares fit the similar results produced by the linear models are clear. Thus, there does not seem to matter which model one uses to get the scores (scale values or true merits) of the treatments, as the scores relation to each other is similar for the three models.

4. Models including ties

Sometimes the assessors are not able to reveal any difference in the pair. This is the case when the treatments are equal in the specific attribute assessed or when the difference is too small to be perceived. The models above do not allow ties, and in such cases the assessors are forced to make a selection and the choice will therefore be made randomly. There is, moreover, a risk that when ties are not allowed and the assessor is unable to make a decision the decision will be based on extraneous criteria that introduce bias into the test. To overcome such effects, ties were sometimes allowed even though no model for handling ties existed. The random answers for comparisons where the assessors could not discern a difference were eliminated most simply by allowing them in the test but ignoring them in the analysis. Some experimenters divided

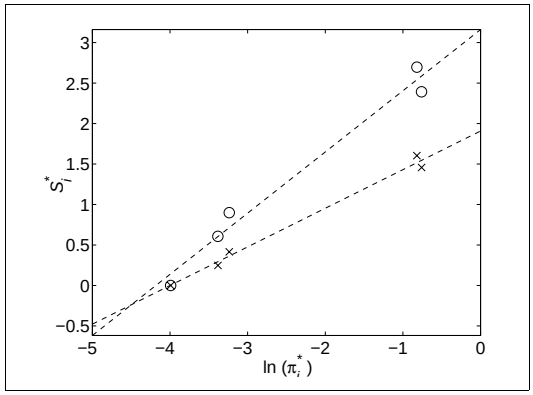

Figure 5. Comparison of scale values. On the horizontal axis is the natural logarithm of the estimated π_i values according to Bradley-Terry. On the vertical axis is the estimated scale value according to Thurstone-Mosteller. Crosses represent Case V and open dots Case III (same Bradley-Terry model for both). Least-squares fit lines are added to show the similarities of the models.

the ties by splitting them. Gridgeman [33] writes about a test made by Hemelrijk where it is proved that leaving ties out of consideration makes a more powerful test than if the ties are equally distributed. However, models that do allow ties in the analysis were developed during the 60's and 70's [11, 12, 5] and there is no longer any need to leave the ties out as information then is lost.

Some who oppose the inclusion of ties highlight the risk of assessors choosing to declare a tie when, with some effort, they might have been able to detect a difference. Gridgeman [33] has investigated this problem and he gave the following recommendations. When discrimination is the objective (i.e. one treatment (sound) is divergent from all the others and it is studied if the divergent treatment is perceived differently) it is better to prohibit ties as the assessors' efficiency of decision making might be offset, but when preference is the objective, ties should be allowed as they add information. In the investigation referred to in section 3.1.1, the number of circular triads was investigated when ties were allowed versus when ties were prohibited [26]. Especially when the sounds were similar, the number of circular triads increased when ties were prohibited. The effect of the possible increased efficiency of the assessors when prohibiting ties might therefore be offset due to circular triads unless only assessors with high consistencies are included, which would also reduce the final number of assessors used.

It has been noticed by the author that, in some cases, assessors' preferences are not normally distributed; in the case of drum sound, some assessors prefer a darker sound while others prefer the more highly pitched sounds. Hence, the sampling distribution seems not to be normal, but binormal, which the models here would not be able to treat correctly. If ties are not permitted, a 50-50 relationship of two treatments would incline one to conclude that the treatments are equal, but if ties are allowed and we still have the same relationship and no ties reported, we can suspect the treatments are not equal but are only given the same amount of preference. When comparing the result in the test example it can be seen that for the C-E pair when no ties were allowed, 13 assessors reported C to be louder than E, and 11 reported E to be louder than C. The same relationship is seen for the A-B pair. However, when ties were allowed, 15 assessors declared a tie for the C-E pair, but only 1 assessor declared a tie for the A-B pair! This information gives the investigators a chance to look further into the problem: are there any differences in the stimuli other than what is asked for that might have influenced the test, and do these differences have different impacts on different groups of assessors? In this test example the character of the sounds of C and E was similar, but A and B had different pitches. In some cases these differences are not considered important, and a mean behaviour is searched for, but in other cases, such as in product development, this could indicate a possible differentiation of the product. Ties also give the proportion of assessors that cannot distinguish between the products, which is also valuable information when deciding to market one of the products or both, or when switching products. Hence, the use of ties provides a tool to give more information about the sound and its influence on the assessors.

4.1. Glenn and David's extension of Thurstone-Mosteller's model

In Thurstone-Mosteller's original model no ties were permitted. A threshold model to handle ties was presented by Glenn and David [11] based on Thurstone's model. If

$$F(\tilde{b}) = \frac{1}{\sqrt{2\pi}} \int_{-\tilde{b}}^{\infty} \mathrm{e}^{-\frac{1}{2}y^2} \mathrm{d}y,$$

the probability $P(T_i \rightarrow T_j), i, j = 1, ..., t$, can, according to Thurstone-Mosteller's model, be written

$$P(T_i \rightarrow T_j) = P(X_i > X_j) = F(S_i - S_j).$$

Glenn and David define an interval of length 2τ around the origin where an assessor cannot distinguish between X_i and X_j but declares a tie. The probabilities that $(T_i \rightarrow T_j)$, $(T_j \rightarrow T_i)$ and $(T_i = T_j)$ are then

$$\begin{aligned} P[(X_i - X_j) > \tau] &= F(-\tau + S_i - S_j) \\ P[(X_i - X_j) < -\tau] &= 1 - F(\tau + S_i - S_j) \\ P[|X_i - X_j| \leq \tau] &= F(\tau + S_i - S_j) - \\ &\quad - F(-\tau + S_i - S_j) \end{aligned}$$

Replacing the parameters above with their estimates gives

$$\begin{aligned} F^{-1}(p_{ij} + p_{0ij}) &= F^{-1}(\tilde{b}_{ij}) = \tau^* + S_i^* - S_j^* \\ F^{-1}(p_{ji} + p_{0ij}) &= F^{-1}(\tilde{b}_{ji}) = \tau^* - S_i^* + S_j^* \end{aligned}$$

and the system is thereafter rearranged so that least square estimates of S_i^* and τ^* can be retrieved. However, Glenn and David show these to be dependent variables using the

normal distribution function and replace the distribution with

$$F(\tilde{b}) = \frac{1}{2}\int_{-\tilde{b}}^{\pi/2} \cos y \, \mathrm{d}y = \frac{1}{2}(1+\sin\tilde{b}).$$

In figure 3 the normal cumulative distribution function (dashed) and the cosine function's cumulative distribution (dotted) are shown. The estimates are then

$$\begin{aligned} \tau^* &= \tfrac{1}{2}\Big[\arcsin(2\tilde{b}_{ij}-1)+\arcsin(2\tilde{b}_{ji}-1)\Big] \\ S_i^* - S_j^* &= \tfrac{1}{2}\Big[\arcsin(2\tilde{b}_{ij}-1)-\arcsin(2\tilde{b}_{ji}-1)\Big] \end{aligned}$$

In a so-called unweighted analysis, corresponding to equal variance of τ and $S_i - S_j$ (Case V), least square estimates of the S_i^* and τ^* are provided. In a following weighted analysis, estimates corresponding to Case III are retrieved by iteration.

4.1.1. Unweighted analysis

The estimate τ^* is given by

$$\tau^* = \frac{1}{t(t-1)}\sum_{i<j} G_{ij}$$

where $G_{ij} = \arcsin(2b_{ij}-1)+\arcsin(2b_{ji}-1)$. S_1^* is set to 0 whereas the other scale values are

$$S_i^* = \frac{1}{t}\left(\sum_{j,j\neq i} H_{ij} + \sum_{i\neq j} H_{ij}\right) \quad \begin{matrix} i = 2,\ldots,t \\ j = 1,\ldots,t \end{matrix}$$

where $H_{ij} = \frac{1}{2}\left[\arcsin(2b_{ij}-1) - \arcsin(2b_{ji}-1)\right]$.

4.1.2. Weighted analysis

The estimates given above are used in the proceeding analysis to get weights, v_{ij} and w_{ij}, to include the effect of various variances of S_i^* and τ.

$$\begin{aligned} v_{ij} &= 1/(1+r_{ij}) \\ w_{ij} &= 1/(1-r_{ij}) \\ r_{ij} &= -\sqrt{\frac{(1-b_{ij}^*)(1-b_{ji}^*)}{b_{ij}^* b_{ji}^*}} \end{aligned}$$

where

$$\begin{aligned} b_{ij}^* &= \tfrac{1}{2}\left[1+\sin(\tau^* + S_i^* - S_j^*)\right] \\ b_{ji}^* &= \tfrac{1}{2}\left[1+\sin(\tau^* - S_i^* + S_j^*)\right] \end{aligned} \tag{12}$$

A new estimate of $\tau^{(k)*}$, $k = 1$ is then

$$\tau^{(k)*} = \sum_{i<j} v_{ij} G_{ij} \Big/ \sum_{i<j} v_{ij} \ .$$

Estimates $S_i^{(k)*}$ are found by

$$\mathbf{S} = \mathbf{W}^{-1}\mathbf{H}_w$$

where

$$\mathbf{S} = \begin{bmatrix} S_2^{(k)*} \\ S_3^{(k)*} \\ \cdots \\ S_t^{(k)*} \end{bmatrix},$$

$$\mathbf{W} = \begin{bmatrix} \sum\limits_{j,j\neq 2} w_{2j} & -w_{23} & \ldots & -w_{2t} \\ -w_{23} & \sum\limits_{j,j\neq 3} w_{3j} & \ldots & -w_{3t} \\ \ldots & \ldots & \ldots & \ldots \\ -w_{2t} & -w_{3t} & \ldots & \sum\limits_{j,j\neq t} w_{tj} \end{bmatrix}$$

and

$$\mathbf{H_w} = \begin{bmatrix} \sum\limits_{j,j\neq 2} w_{2j}H_{2j} \\ \sum\limits_{j,j\neq 3} w_{3j}H_{3j} \\ \cdots \\ \sum\limits_{j,j\neq t} w_{tj}H_{tj} \end{bmatrix}$$

With these estimates, new estimates $S_i^{(k)*}$ and $\tau^{(k)*}$, $k = 2, ..., k_{tol}$ are calculated until the difference between each iteration is less than the requested tolerance. If $n_{ij} \neq n$, that is if an unbalanced design is used, the weights are changed to $v_{ij} = n_{ij}/(1+r_{ij})$ and $w_{ij} = n_{ij}/(1-r_{ij})$ and the first estimates are given by setting $v_{ij} = w_{ij} = n_{ij}$.

This model is described briefly in [34], where another model for handling ties, denoted the uniform model, is presented as well. However, it will not be treated here.

Goodness to fit

To check the validity of the model Glenn and David present a χ^2 test of goodness of fit.

$$\chi^2 = \sum_{i<j}\frac{1}{n_{ij}}\left[\frac{(a_{ij}-n_{ij}p_{ij}^*)^2}{p_{ij}^*} + \right.$$

$$\left. + \frac{(a_{ji}-n_{ij}p_{ji}^*)^2}{p_{ji}^*} + \frac{(a_{0ij}-n_{ij}p_{0ij}^*)^2}{p_{0ij}^*}\right] \tag{13}$$

p_{ij}^*, p_{ji}^* and p_{0ij}^* are estimates calculated as

$$\begin{aligned} p_{ij}^* &= (1-b_{ji}^{(k)*}) \\ p_{ji}^* &= (1-b_{ij}^{(k)*}) \\ p_{0ij}^* &= (b_{ij}^{(k)*} + b_{ji}^{(k)*} - 1) \end{aligned}$$

where $b_{ij}^{(k)*}$ and $b_{ji}^{(k)*}$ are calculated as in equation (12) but with the new estimates. a_{ij}, a_{ji} and a_{0ij} are the number of times ($T_i \rightarrow T_j$), ($T_j \rightarrow T_i$) and ($T_i = T_j$) respectively. For large samples χ^2 has a chi-square distribution with $t(t-2)$ degrees of freedom if the model is valid. (There are $t(t-1)/2$ pairs yielding two independent observations each and t estimates.) Pairs with $n_{ij} = 0$ must be omitted

and the degrees of freedom reduced by two. Small numbers of n_{ij},p^*_{ij},p^*_{ji} or p^*_{0ij} can distort the test statistic. As a consequence, it is here suggested that terms where $n_{ij}p^*_{ij}$, $n_{ij}p^*_{ji}$ or $n_{ij}p^*_{0ij}$ is less than 1 be omitted. If only one term (out of three) for a pair is omitted, delete one degree of freedom; otherwise all terms of the pair should be omitted and the degrees of freedom be reduced by two.

4.2. Rao and Kupper's extension of the Bradley-Terry model

The idea to the model by Rao-Kupper [5] is similar to the one by Glenn-David [11]; when the difference between two treatments is smaller than a certain value, or threshold, the assessors will declare a tie. Glenn and David apply this idea to the model by Thurstone-Mosteller whereas Rao and Kupper apply it to the model by Bradley-Terry. The probability of choosing T_i when compared to T_j, $\pi_{i,ij}$, is then

$$P(T_i{\rightarrow}T_j)=\frac{1}{4}\int_{-(\ln\pi_i-\ln\pi_j)+\eta}^{\infty}\operatorname{sech}^2(y/2)\mathrm{d}y=$$

$$=\frac{\pi_i}{\pi_i+\theta\pi_j} \qquad (14)$$

where η=ln(θ) is the sensory threshold for the assessor, cf eq. (1). The probability for a tie, $\pi_{0,ij}$, is

$$P(T_i{=}T_j)=\frac{1}{4}\int_{-(\ln\pi_i-\ln\pi_j)-\eta}^{(\ln\pi_i-\ln\pi_j)+\eta}\operatorname{sech}^2(y/2)\mathrm{d}y=$$

$$=\frac{\pi_i\pi_j(\theta^2-1)}{(\pi_i+\theta\pi_j)(\theta\pi_i+\pi_j)}.$$

When the sensory threshold is set to be constant it is assumed to be the same for all participating assessors during the entire test. Although not commented on further here, groups with various sensory thresholds are treated in [5].

Estimates of the t treatment ratings and the sensory threshold, are then found by the use of maximum likelihood functions as in the Bradley-Terry model. The estimates, π^*_i of π_i and θ^* of $\theta = \mathrm{e}^{\eta}$, are given by an iterative process with the constraint $\sum_i^t \pi_i = 1$. Starting values may be $\pi_i^{(0)^*} = 1/t$ and $\theta^* = 2N/a - 1$ where $N=\sum_{i<j} n_{ij}$, $n_{ij} = a_{ij} + a_{ji} + a_{0ij}$, is the total number of comparisons made for the pair (i,j).[4] a_{ij}, a_{ji} and a_{0ij} are the number of times $(T_i \rightarrow T_j)$, $(T_i \rightarrow T_j)$ and $(T_i = T_j)$ respectively. $a=\sum_{i<j}(a_{ij}+a_{ji})$.

$$\hat{\pi}_i^{(k+1)}=\frac{b_i}{\sum\limits_{j,j\neq i}\left(\frac{b_{ij}}{\pi_i^{(k)^*}+\theta^{(k)^*}\pi_j^{(k)^*}}+\frac{b_{ji}\theta^{(k)^*}}{\theta^{(k)^*}\pi_i^{(k)^*}+\pi_j^{(k)^*}}\right)}$$

$$\pi_i^{(k+1)^*}=\hat{\pi}_i^{(k+1)}\Big/\sum_i\hat{\pi}_i^{(k+1)}$$

$$\theta^{(k+1)^*}=1+\frac{\theta^{(k)^*}}{\theta^{(k)^*}+1}\left[\frac{2(N-n)}{\sum\limits_{i\neq j}b_{ij}\pi_j^{(k+1)^*}\left(\pi_i^{(k+1)^*}+\theta^{(k)^*}\pi_j^{(k+1)^*}\right)^{-1}}\right]$$

$b_i=\sum_{j,j\neq i} b_{ij}$, and b_{ij} is the number of times that either treatment T_i is chosen over T_j or a tie is declared.

Goodness to fit

The appropriateness of the model is checked with the same χ^2 test as in the model by Glenn-David, equation (13).

4.3. Davidson's extension of the Bradley-Terry model

Davidson presented some years after Rao and Kupper another extension to handle ties based on the Bradley-Terry model. His approach is to ensure that Luce's [35] "axiom of choice" is fulfilled. According to Luce there are alternatives that should be irrelevant to the choice, ensured to be irrelevant when $P(T_i \rightarrow T_j)/P(T_j \rightarrow T_i) = \pi_i/\pi_j$, which the model by Rao-Kupper does not fulfil. Davidson assumes that $P(T_i = T_j) = \nu\sqrt{P(T_i{\rightarrow}T_j)P(T_j{\rightarrow}T_i)}$ where ν is seen as an index of discrimination. The assumption of a geometric mean is based on the fact that the merits, lnπ, can be represented on a linear scale. The probability to choose T_i when presented with T_j then becomes

$$P(T_i{\rightarrow}T_j) = \frac{\pi_i}{\pi_i{+}\pi_j{+}\nu\sqrt{\pi_i\pi_j}}$$

$$P(T_i{=}T_j) = \frac{\nu\sqrt{\pi_i\pi_j}}{\pi_i{+}\pi_j{+}\nu\sqrt{\pi_i\pi_j}}.$$

The estimates are obtained in similar ways as in the Bradley-Terry model, using maximum likelihood functions:

$$\hat{\pi}_i^{(k+1)}=s_i\Big/\sum_{j,j\neq i}\frac{n_{ij}\left(2+\nu^{(k)^*}\sqrt{\pi_j^{(k)^*}/\pi_i^{(k)^*}}\right)}{\pi_i^{(k)^*}+\pi_j^{(k)^*}+\nu^{(k)^*}\sqrt{\pi_i^{(k)^*}\pi_j^{(k)^*}}}$$

$$\pi_i^{(k+1)^*}=\hat{\pi}_i^{(k+1)}\Big/\sum_i\hat{\pi}_i^{(k+1)}$$

$$\nu^{(k+1)^*}=a_0\Big/\sum_{i<j}\frac{n_{ij}\nu^{(k)^*}\sqrt{\pi_j^{(k+1)^*}\pi_i^{(k+1)^*}}}{\pi_i^{(k+1)^*}+\pi_j^{(k+1)^*}+\nu^{(k)^*}\sqrt{\pi_i^{(k+1)^*}\pi_j^{(k+1)^*}}}$$

where $s_i = 2a_i + a_{0i}$, $i = 1,...,t$, $a_i = \sum_j a_{ij}$ and $a_{0i} = \sum_j a_{0ij}$ where a_{ij} is the number of times T_i was chosen over T_j and a_{0ij} is the number of ties when the pair (i,j) is compared. $n_{ij} = a_{ij} + a_{ji} + a_{0ij}$, that is the

[4] An extended version of the model in [5] is here given to allow unbalanced designs. In [5] n_{ij} is assumed to be constant for all pairs and $N = nt(t-1)$.

Table VII. Scale values, true merits and goodness to fit P-value for the drum sounds, DS, using the Glenn-David (GD), Rao-Kupper (RK) and Davidson (D) models. The threshold values for the different models, τ^*, η^*=ln(θ^*), ν^* are given in the row 'Th'.

DS	GD	RK	D	DS	GD	RK	D
A	0	-0.51	-0.51	A_3	0	-2.75	-3.75
B	0.00	-1.24	-1.06	B_3	-0.11	-3.03	-4.16
C	-0.81	-2.97	-3.62	C_3	-0.62	-4.28	-6.19
D	-1.06	-3.59	-4.50	D_3	-1.18	-6.74	-9.97
E	-0.91	-3.35	-4.12	E_3	0.07	-2.32	-3.20
				F_3	-1.49	-8.40	-12.46
				G_3	0.64	-0.26	-0.09
Th	0.36	0.95	1.30		0.07	1.21	1.83
P	0	0	0		0	0.002	0.15

total number of comparisons of pair (i,j). a_0 is the total number of ties, $a_0 = \sum_{i<j} a_{0ij}$.

Davidson [12] and Bradley [9] notice that the models by Davidson and Rao-Kupper are asymptotically equal and choosing which method to use is a matter of which of the ideas seems more appealing to the experimenter. David [10] points out that the fulfilment of the Luce axiom might not be required. Both models have the property that the probability of a tie is highest when $\pi_i/\pi_j = 1$. However, the model by Davidson has the properties to give a ranking that is always consistent with a ranking common in sports, where a win is awarded 2 points and a tie 1 point; the Rao-Kupper model does not have these properties.

Goodness to fit

The appropriateness of the model is checked with the same χ^2 test as in the model by Glenn-David, equation (13).

4.4. Examples using tie models

The result using these three models on the data from examples 2 and 3 in tables II and III are shown in table VII. The scale values, S_i^*, and true merits, ln(π_i), are calculated until the relative change from one iteration to the next is less than 0.001. The goodness to fit data were calculated omitting terms where $n_{ij}p_{ij}^*$, $n_{ij}p_{ji}^*$ or $n_{ij}p_{0ij}^*$ is less than 1; otherwise all P-values would be less than 0.0001. The omittance of these terms in the χ^2 test can be justified by looking at, as an example, the term $(a_{ij} - n_{ij}p_{ij}^*)^2/n_{ij}p_{ij}^*$ from the comparison D_3–E_3 using the model by Davidson. As one assessor out of 31 in the test chose D_3 over E_3 and the model predicts 0.0001, the term in the χ^2 test becomes equal to 7659 which is 97% of the total sum if all terms are included.

The P-values for example 2 are all so low that the estimates p_{ij}^* are not equal to p_{ij}. The main reason for the low values are the pair $A-B$. As discussed in the beginning of section 4 the number of assessors that find A louder than B is 14, 9 assessors find B louder than A, but only 1 declares a tie. None of these models can predict this behaviour. In cases like this the poor goodness to fit can help the investigator to find these deviations so that a deeper analysis can be made; the poor fit might be due to a specific pair whose differences then should be investigated; maybe a multivariate analysis should be used where several characteristics in the treatments are investigated. When no tie was permitted the prediction was good and this behaviour was hidden and therefore information was lost.

For example 3 the P-values for both the Glenn-David and Rao-Kupper models indicate that the estimates are not good enough. The model by Glenn-David fails to predict the threshold correctly and the number of ties is too low. As the idea of Rao-Kupper and Glenn-David is similar but applied on different distributions and solution techniques it is not surprising, considering what Stern found, that the performance of the Rao-Kupper model is somewhere in between the performance of the Glenn-David and Davidson models. The predictions of p_{ij} using the Rao-Kupper and Davidson models are similar when looking at them without performing the goodness to fit test, but the goodness to fit test indicates the Davidson model is the better one. However, this is not the general result comparing these models; in [4], the goodness to fit was better using the model by Rao-Kupper (P-value $= 0.08 > 0.01$); however the model by Glenn-David gave the least fit (P-value$< 10^{-12}$) there as well.

In figure 6, the scale values and true merits are plotted comparing two models at a time. Least-squares fit lines are also added. The data fit well to the least-square fit although the models have different P-values. If the objective of the test is to rank or get scale values or true merits of the treatments, the choice of model does not seem to be crucial. If predictions of p_{ij} are to be used in a further analysis, the models by Rao-Kupper and Davidson seem more appropriate; preferably both models should be tested and the model with the better goodness to fit used.

In figure 7 the scale values and true merits of the treatment in examples 1 and 2 are plotted against each other. Although they follow the least-squares fit nicely, a difference in ranking can be noticed; A and B switch ranking. As pointed out above, the models that have been used here cannot handle the situation when similar numbers of assessors declare A to be louder than B, and B to be louder than A, but still no assessors declare a tie. Another effect that has to be considered when discussing the reasons for the different ranking is the confidence interval of the estimates. Are there any significant differences between the true merits? In the following section this question will be addressed.

5. Estimation of differences between sounds

The most sensitive comparison to estimate if the stimuli or treatments are different or similar is made using only two stimuli. Procedures to perform such tests are presented in an ISO-standard prepared for food products [36]. The method used there can also be found in [10]. The method

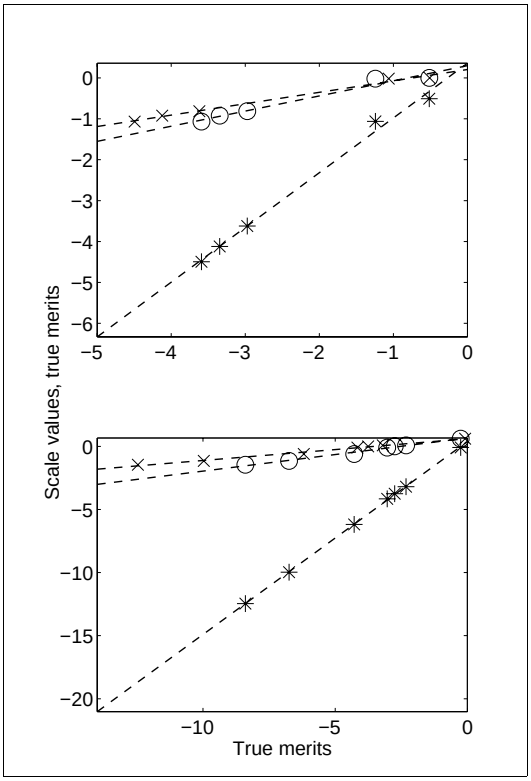

Figure 6. Comparison of scale values and true merits from examples 2 (top) and 3 (bottom). The result using the Rao-Kupper model on the horizontal axis and the model by Glenn-David and Davidson on the vertical axis are shown by the circles and stars respectively. The crosses represent the result using the model by Davidson on the horizontal axis and the model by Glenn-David on the vertical axis. Least-squares fit lines are added to show the similarities of the models.

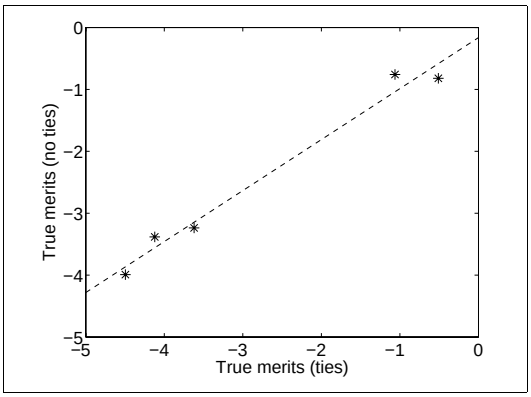

Figure 7. True merits when allowing ties (Davidson model, horizontal axis) compared with true merits when ties are prohibited (Bradley-Terry model, vertical axis). The line represents a least-squares fit.

is useful in investigations of whether a product's sound is affected by a specific change such as a new distributor of an engine part, etc. However, in many listening tests more than two sound stimuli are involved, and for those cases, methods to estimate differences between two stimuli are presented in the following sections.

5.1. Confidence intervals

A simple way to check whether there are any significant differences between the scale values of true merits is to compare their confidence intervals. The null hypothesis that a pair is equal can be rejected at the chosen significance level if the confidence intervals do not overlap. If it is decided prior to performing the listening tests to look at the results in which only the pair $A-B$ is to be compared, then the chosen significance level, e.g. α=0.05, is used to calculate the confidence intervals for the two treatments. However, in many cases it is only after performing the test that it fist becomes clear which treatments that seem equal, and often several pairs are compared. When performing several, k, significance tests each at the α level, the probability of making at least one rejection of the null hypothesis inappropriately (Type I error [25]) is $1-(1-\alpha)^k$. As an example; if k=3 and α=0.05, the effective significance level for Type I error is 0.14 instead of 0.05. A common way to deal with this problem is to use the Bonferroni method where α is replaced with α/t, where t is the number of treatments in the whole listening test.

The limits of the confidence intervals for the scale values, S_i, $i=1,...,t$ of the Thurstone-Mosteller model are retrieved for Case V as

$$\left(S_i^* - z_{\alpha/2}\frac{\sqrt{t-1}}{t}, S_i^* + z_{\alpha/2}\frac{\sqrt{t-1}}{t}\right)$$

$z_{\alpha/2}$ is the normal deviate at the chosen significance level α, $S_i^* = 1/t\sum_{j\neq i} z_{ij}$. The confidence interval limits for Case III are

$$\left(S_i^* \pm \hat{t}_{\alpha/2,(t-1)}\frac{\sqrt{\sigma_i^2(t-2)+\sum\sigma_j^2}}{t}\right)$$

where $\hat{t}_{\alpha/2,(t-1)}$ is the Student's t-distribution at the chosen significance level with $(t-1)$ degrees of freedom. $S_i^* = 1/t\sum_{j\neq i}(z_{ij}\sqrt{\sigma_i^2+\sigma_j^2})$. A drawback of the Bonferroni method is that even though the risk of incorrectly producing a difference on an individual test is reduced, the risk of making the Type II error is increased, i.e. no difference is declared even though there in fact is a difference. When Case V applies (independent treatment and equal variances), analysis of variance, ANOVA, can be used on the matrix of normal deviates to test if any of the treatments is divergent. Thereafter, a multiple comparison procedure such as the Tukey test can be used to assess which scale values are different [25].

An overall hypothesis test $\pi_i = ... = \pi_t$ for the Bradley-Terry model is presented in [8, 9]. It is recommended to perform this test especially when all treatments

are similar to reduce the risk of making a Type II error before continuing the work. To receive confidence intervals for the Bradley-Terry model the dispersion matrix $\mathbf{\Sigma} = [\sigma_{ij}]$, also called the covariance matrix, needs to be calculated. An estimate $\mathbf{\Sigma}^*$ of $\mathbf{\Sigma}$ is retrieved as

$$\sigma_{ij}^* = \text{cofactor of } \lambda_{ij}^* \text{ in } \begin{bmatrix} \mathbf{\Lambda}^* & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix} \Big/ \begin{vmatrix} \mathbf{\Lambda}^* & \mathbf{1} \\ \mathbf{1}' & 0 \end{vmatrix}$$

where $\mathbf{\Lambda}^* = [\lambda_{ij}^*]$, $\mathbf{1}'$ is the t-dimensional unit row vector and

$$\lambda_{ii}^* = \frac{1}{\pi_i^*} \sum_{j,j \neq i} \mu_{ij} \pi_j^* / (\pi_i^* + \pi_j^*)^2, \quad i = 1, ..., t$$

$$\lambda_{ij}^* = -\mu_{ij} / (\pi_i^* + \pi_j^*)^2, \quad i \neq j, \quad i, j = 1, ..., t$$

where $\mu = n_{ij}/N$. As Bradley shows $\pi_i^* \sqrt{N}(\ln(\pi_i^*) - \ln(\pi_i))/\sqrt{\sigma_{ij}^*}$ to be standard normal for large N [37] the confidence interval for $\log(\pi_i)$ is

$$\left(\ln \pi_i^* - z_{\alpha/2} \sqrt{\frac{\sigma_{ii}^*}{(\pi_i^*)^2 N}}, \ln \pi_i^* + z_{\alpha/2} \sqrt{\frac{\sigma_{ii}^*}{(\pi_i^*)^2 N}} \right)$$

where $z_{\alpha/2}$ is the normal deviate at the chosen significance level α. α is replaced with α/t when the Bonferroni method is applied. The dispersion matrices for the Rao-Kupper and Davidson models are retrieved in a similar way although they are not shown here [5, 12].

5.1.1. Examples

In figure 8 the 95% confidence intervals for the test data in example 1 using the Bradley-Terry and Thurstone-Mosteller Case V models are shown. The Bonferroni method has been applied; α=0.01. The widths of the confidence intervals of the three models should not be compared directly with each other but in proportion with their scale values/true merits. Still, it is seen that the confidence intervals using the Thurstone-Mosteller model are larger than those of the Bradley-Terry model; all intervals overlap using the former model, which is not the case for the latter. The large difference in the confidence intervals can be explained by how the variances are estimated. The solution in [28] was developed when iterative solutions were avoided. The estimates would probably be more efficient if the solution of the Thurstone-Mosteller model is based on the maximum likelihood method as in the Bradley-Terry model.

In figure 9 the 95% confidence intervals for the test data in example 2 using the Rao-Kupper and Davidson models is shown. The Bonferroni method has been applied; α=0.01. A and B are not equal on the 0.05 level to any other stimuli as their confidence intervals do not overlap the others. C-D-E's confidence intervals overlap in the Rao-Kupper model and the hypothesis that they are equal cannot be rejected on the chosen significance level. A similar result is given by the Davidson model.

Comparing the result without ties using the Bradley-Terry model and with ties using the Rao-Kupper or Davidson model, it can be seen that less overlap of the confidence intervals is achieved when ties are allowed. Please note that the widths of the models' confidence intervals should not be compared directly with each other in absolute numbers but in proportion with their scale values/true merits. The reversed ranking of A and B in the test with and without ties could be explained by the overlapping confidence intervals in the test without ties. However, as discussed earlier, the paired comparison models presented here might not be suitable since the expected (predicted) number of ties for A and B does not fit the seemingly binormal sampling distribution of the raw data.

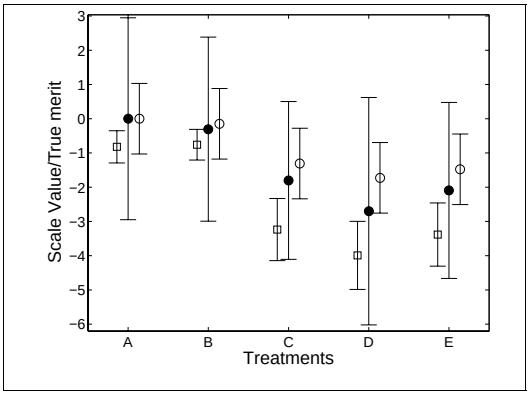

Figure 8. Scores (scale values/true merits) and their 95% confidence intervals for the test data in example 1 using the models of Bradley-Terry (squares), Thurstone-Mosteller Case III (filled circles) and Case V (open circles).

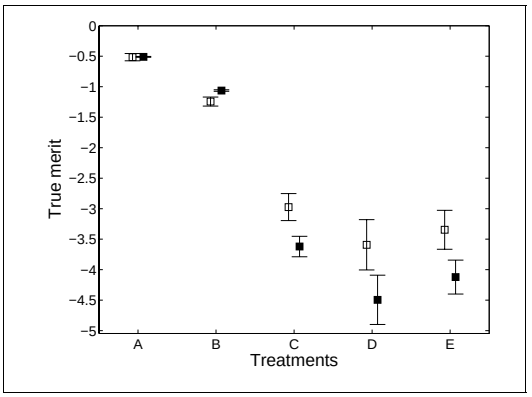

Figure 9. Scores (true merits) and their 95% confidence intervals for the test data in example 2 using the models of Rao-Kupper (open squares) and Davidson (filled squares).

5.2. Confidence regions

The method of using confidence intervals is, however, rough as no consideration of any dependent behaviour of the treatments has been made; if one of the treatments is changed the treatment ratings for all the other treatments

will change in the Bradley-Terry model. This is the major reason why ANOVA-tests are not appropriate; the other reason is that in an ANOVA-test it is also assumed that the variances are equal, which they are not in many listening tests. By the use of the dispersion matrix, confidence regions are created in which the dependent behaviour is included, enabling tests of equality. The (1-α) confidence region for t parameters is an ellipsoidal region for which

$$N(\ln\Pi - \ln\Pi^*)'\mathbf{D}\boldsymbol{\Sigma}^{*-1}\mathbf{D}(\ln\Pi - \ln\Pi^*) \leq \chi^2_{\alpha,t}$$

where $\Pi^* = [\pi_1^*, ..., \pi_t^*]$, is the estimates of $\Pi = [\pi_1, ..., \pi_t]$, $\boldsymbol{\Sigma}^*$ is the same dispersion matrix as above and $\mathbf{D}$ is a diagonal matrix with elements $1/\pi_i^*$ [9]. The confidence region when comparing the two treatments T_i and T_j is then governed by

$$\begin{aligned}\chi^2_{\alpha,t} \geq \frac{N}{\sigma_{ii}\sigma_{jj} - \sigma_{ij}^2}((\ln\pi_i - \ln\pi_i^*)^2\sigma_{jj}\pi_i^{*2} - \\ -(\ln\pi_i - \ln\pi_i^*)(\ln\pi_j - \ln\pi_j^*)2\sigma_{ij}\pi_i^*\pi_j^* + \\ +(\ln\pi_j - \ln\pi_j^*)^2\sigma_{ii}\pi_j^{*2})\end{aligned} \quad (15)$$

The corresponding P-value, that is the lowest significance level that would lead to a rejection of the hypothesis $\ln\pi_i = \ln\pi_j$, is given by calculating the location where the confidence region first cuts the plane of symmetry for decreasing values of α. The location is given by

$$\ln\pi_i = \ln\pi_j = \frac{\ln\frac{(\pi_i^*)^{\sigma_{jj}^*(\pi_i^*)^2}(\pi_j^*)^{\sigma_{ii}^*(\pi_j^*)^2}}{(\pi_i^*\pi_j^*)^{\sigma_{ij}^*\pi_i^*\pi_j^*}}}{\sigma_{jj}^*(\pi_i^*)^2 - 2\sigma_{ij}^*\pi_i^*\pi_j^* + \sigma_{ii}^*(\pi_j^*)^2}. \quad (16)$$

The location $\ln\pi_i = \ln\pi_j$ is thereafter used as input data in equation (15) to retrieve the P-value.

In figure 10 the confidence region for two treatments at two significance levels are shown; ellipsoidal regions are received. The estimates of $\ln\pi_i$ and $\ln\pi_j$ should exist somewhere in that region at the significance level chosen. A hypothesis test of any point outside the region will be rejected on the chosen significance level. If the region is crossed by the plane of symmetry the hypothesis that they are equal cannot be rejected. As the smaller region in figure 10 does not cross the plane of symmetry the treatments are different at that level.

General conclusions regarding the estimates' dependency and their variance can be drawn based on the confidence region if either the values on the vertical axis of figure 10 are multiplied with π_j^*/π_i^* as in figure 11 or if the confidence region of the estimates π is drawn. By looking at the region we can check whether our estimates are independent and if the variances are homogeneous. If the region forms a circle, the variance is the same for the estimates and the estimates are independent. If an ellipse is given and if its axes are parallel to the axes of the treatment ratings, the estimates are independent but their variances are not homogeneous. If the ellipse is not parallel to any of the axes, like the one in Figure 10, the estimates are not independent. Furthermore, if it is making the angle ($\pi/4$) with the horizontal axis, the estimates are not independent but their variances are homogeneous.

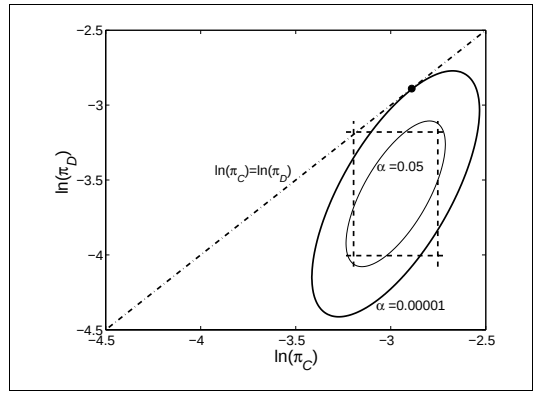

Figure 10. Confidence region for treatments C and D from example 2 using the Rao-Kupper model at two significance levels. The 95% confidence intervals are shown with the dashed lines. The location where the confidence region first cuts the plane of symmetry for decreasing values of α given by equation (16) is shown with a dot.

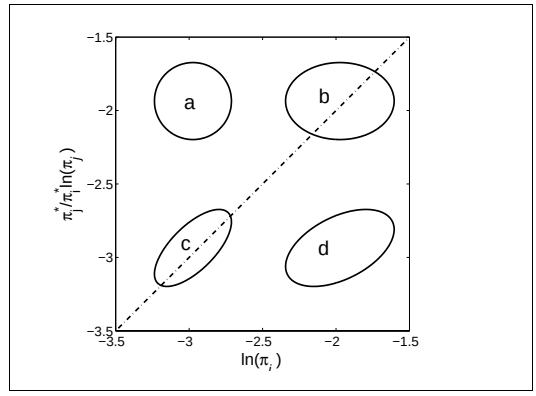

Figure 11. The four regions illustrate: a) independent estimates and equal variance, b) independent estimates and unequal variances, c) dependent estimates and equal variance, d) dependent estimates and unequal variances.

A further refinement of the procedure can be made using contrasts [38, 9] but it will not be described in this paper.

5.2.1. Example

In figure 10 the confidence region for treatments C and D from example 2 using the Rao-Kupper model is plotted at two significance levels. The P-value is 0.00001. As a comparison, the 95% confidence intervals are plotted with the dashed lines; since the area they create crosses the symmetry line (dash-dotted), any hypothesis that they are equal cannot be rejected using confidence intervals. However, in the figure the 95% confidence region is plotted with the thinner solid line so they are in fact not equal on

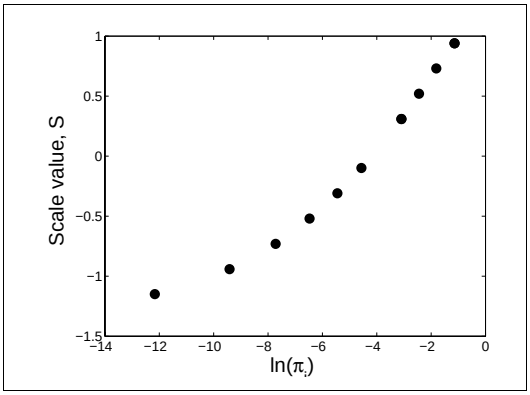

Figure 12. Comparison of test result using model by Ando and Singh and Bradley-Terry model for data from one subject comparing every pair only once.

the $\alpha = 0.05$ level. (Both the interval and the region are calculated using the Bonferroni method.)

6. Miscellaneous

For cases where only one subject has compared the stimuli once, Thurstone's model in its original form is unable to produce scale values of the stimuli. Ando and Singh presented a model in [39] where the cumulative distribution function of the normal distribution is expended in a Taylor series where only first-order terms are considered. The sum in equation 5 can then be made on the number of selections before translating the value by the use of the Taylor series to a scale value. This approximation will work best when no extremes are considered. The same situation can, however, be treated with the Bradley-Terry model. The data presented in [39] were used to produce the plot in figure 12. On the horizontal scale are the natural logarithms of the π_i and on the vertical scale are the scale values given in [39]. Even though the given data are limited, the two models here produce the same ranking. The scale values are not proportional as there is no straight line, but this is not surprising as the measure in [39] is linearized and the maximum likelihood estimation is based on data that are either one or zero.

As mentioned in the introduction, a drawback of the method of paired comparison is the rapid growth of comparisons required when the number of treatments increases. An approach to reduce the number of comparisons is to make a balanced incomplete design, in which some comparisons are completely omitted. Kendall presents two minimum requirements for an incomplete design: the selected treatments should appear equally often and it must not be possible to divide the treatments into subsets where no comparisons are made between the subsets [40]. The procedure is explained in [10] and tables that enable balanced incomplete designs can be found in [32, 41]. If, prior to the test, specific interest lies in any specific pair, make sure the pair is included as the confidence interval for those pairs excluded increases.

Several extensions and variations to the paired comparison model exist; for example in [42] a method to include effects of time-varying data is presented. In cases where more than a three-point scale ($T_i \rightarrow T_j$, $T_i = T_j$, $T_i \leftarrow T_j$) is requested the method by Scheffé [43] is recommended. It does not only provide a means of analysing data based on a 7- or 9-point scale but within-pair order effects can also be investigated. Many aspects that have not been dealt with here, such as within-pair order effects, circular triads, consistency tests, triple comparisons and multivariate paired comparisons where several characteristics in the treatments are investigated, are addressed in [10]. A bibliography on the method of paired comparisons can be found in [44].

When analysing the result it is recommended to check the distribution of the answers to see if there is any reason to doubt that the question has been clear to the subjects. No comments regarding the choice of question have been made; naturally it is of great importance to consider the choice thoroughly.

7. Conclusion and recommendations

Mosteller points out in [7] that discrepancies from the normality assumption are not important to the method of paired comparisons. This statement is indirectly confirmed by Stern [14], and it is seen in figures 5 and 6 as well, that the paired comparison analysis is not very sensitive to the choice of distribution within the class of linear models.

The recommendation from Gridgeman [33] still seems to hold: When discrimination is the objective (one treatment is different from all the others) it is better to prohibit ties as the assessors' efficiency of decision making might be offset, but when preference is the objective, ties should be allowed as they add information. In the test example the confidence intervals of the treatment ratings are shorter when ties are allowed. The use of ties enabled the discovery that the sampling distribution of the normal for one pair of the sounds is binormal; a test without ties would conceal that information.

By performing goodness to fit tests an indication of the appropriateness of the model can be given. If poor goodness to fit is achieved, the reason should be investigated before conducting further tests.

The following recommendations are therefore given dependent on the objective of the investigations and character of the sound:

- If the objective is to determine whether or not there is a perceptible difference between two (and only two) sound stimuli, consider the ISO-standard [36].
- If the objective is to get a ranking of the treatments, and a complete balanced design is used, base the ranking on the number of wins for each treatment. If ties are allowed a win is awarded 2 points and a tie 1 point. If the design is incomplete, use the Bradley-Terry model as

its maximum likelihood solution better uses all available data than the solution presented in section 3.1.1. If ties are allowed use either the model by Rao-Kupper or Davidson's model.

- If the objective is to achieve scale values and no ties are allowed, use the model by Thurstone-Mosteller. If the sounds that are to be compared are homogeneous, hence not varying in many different properties such as loudness, sharpness, tonality etc., the variances are more likely to be equal and Case V is more likely to be applicable. If ties are allowed use either the model by Rao-Kupper or Davidson's model.
- If estimations of differences are to be made and no ties are allowed, use the Bradley-Terry model and calculate the confidence regions. If ties are allowed, base the choice of either the Rao-Kupper or Davidson model on which model gives the best goodness to fit before calculating the confidence regions.

Acknowledgement

The authors wish to thank Professor Björn Holmquist at the Department of Statistics and Tobias Rydén at the Department of Mathematical Statistics at Lund University for their suggestions and valuable discussions. 'The Building and Its Indoor Environment' research school at Lund University is thankfully acknowledged for the financial support.

References

[1] R. Guski: Psychological methods for evaluating sound quality and assessing acoustic information. Acta Acustica united with Acustica **83** (1997) 765–774.

[2] M. Khan, O. Johansson, U. Sundbäck: Evaluation of annoyance reponse to engine sounds using different rating methods. Proceedings of InterNoise 96, Liverpool, UK, 1996. 2517–2520.

[3] E. Parizet, N. Hamzaoui, G. Sabatié: Comparison of some listening test methods: A case study. Acta Acustica united with Acustica **91** (2005) 356–364.

[4] A.-C. Johansson, P. Hammer, E. Nilsson: Prediction of subjective response from objective measurements applied to walking sound. Acta Acustica united with Acustica **90** (2004) 161–170.

[5] P. Rao, L. Kupper: Ties in paired-comparison experiments: a generalization of the Bradley-Terry model. Journal of the American Statistical Association **62** (1967) 194–204 Corrigenda 63, p.1550.

[6] L. Thurstone: Psychophysical analysis. The American journal of psychology **38** (1927) 368–389.

[7] F. Mosteller: Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. Psychometrika **16** (1951) 3–9, 203–206, 207–218.

[8] R. Bradley, M. Terry: The rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika **39** (1952) 324–345.

[9] R. Bradley: Paired comparisons: Some basic procedures and examples. – In: Handbook of Statistics Vol. 4. Elsevier Science Publisher, North-Holland Publishing Co, Amsterdam, 1984, 299–326.

[10] H. David: The method of paired comparisons, 2nd ed. Griffin, London, 1988.

[11] W. Glenn, H. David: Ties in paired-comparison experiments using a modified Thurstone-Mosteller model. Biometrics **16** (1960) 86–109.

[12] R. Davidson: On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. Journal of the American Statistical Association **65** (1970) 317–328.

[13] H. Stern: A continuum of paired comparisons models. Biometrika **77** (1990) 265–273.

[14] H. Stern: Are all linear paired comparison models empirically equivalent? Mathematical Social Sciences **23** (1992) 103–17.

[15] A.-C. Johansson, E. Nilsson, P. Hammer: Evaluation of drum sound with ISO tapping machine. Building Acoustics **12(2)** (2005) in press.

[16] A.-C. Johansson, E. Nilsson: Measurement of drum sounds 1636-03 (04207). Technical Report NT573, Nordic Innovation Centre, Oslo, 2005.

[17] G. Fechner: Elemente de psychophysik. Breitkopf und Härtel, Leipzig, 1860.

[18] G. Fechner: Elements of psychophysics 1. Trans. by H.E. Adler, Holt, Rinehart and Winston, New York, 1965.

[19] W. Torgerson: Theory and methods of scaling. Wiley, New York, 1958.

[20] L. Thurstone: A law of comparative judgement. Psychological Review **34** (1927) 273–286.

[21] L. Thurstone: The method of paired comparisons for social values. Journal of Abnormal and Social Psychology **21** (1927) 384–400.

[22] Y. Takane: Maximum likelihood estimation in the generalized case of Thurstone's model of comparative judgment. Japanese Psychol. Res. **22** (1980) 186–196.

[23] A. L. Edwards: Techniques of attitude scale construction. New York, Appleton-Century-Crofts, Inc., 1957.

[24] J. P. Guilford: Psychometric methods. McGraw-Hill, New York, 1954, 154–177.

[25] D. Montgomery: Design and analysis of experiments, 5th ed. John Wiley and Sons, New York, 2001.

[26] E. Parizet: Paired comparison listening tests and circular error rates. Acta Acustica united with Acustica **88** (2002) 594–598.

[27] D. Mao, W. Yu, Y. Gao, Z. Wang: Error rating of paired comparison listening test for sound quality assessment. Proceedings of ICA 2004, 2004. 1799–1802.

[28] R. Burros: The application of the method of paired comparisons to the study of reaction potential. Psychological Review **58** (1951) 60–66.

[29] U. Böckenholt, R.-C. Tsai: Individual differences in paired comparison data. British Journal of Mathematical and Statistical Psychology **54** (2001) 265–277.

[30] E. Zermelo: Die Berechnung der Turnier-Ergebnisse als ein Maximum-problem der Wahrscheinlichkeitsrechnung. Meth. Zeit. **29** (1929) 436–460.

[31] L. Ford: Solution of a ranking problem from binary comparisons. Amer. Math. Monthly **64** (1957) 28–33.

[32] O. Dykstra: Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs. Biometrics **16** (1960) 176–188.

[33] N. Gridgeman: Pair comparison, with and without ties. Biometrics **15** (1959) 382–388.

[34] W. Batchelder, N. J. Bershad: The statistical analysis of a thurstonian model for rating chess players. Journal of Mathematical Psychology **19** (1979) 39–60.

[35] R. Luce: Individual choice behavior. Wiley, New York, 1959.

[36] ISO/DIS 5495, Sensory analysis — Methodology — Paired comparison test. Revision of second edition ISO5495:1983, 2004.

[37] R. Bradley: Rank analysis of incomplete block designs. III. Some large-sample and results on estimation and power for a method of paired comparisons. Biometrika **42** (1955) 450–470.
[38] R. Bradley, A. El-Helbawy: Treatment contrasts in paired comparisons: Basic procedures with application to factorials. Biometrika **63** (1976) 255–262.
[39] Y. Ando, P. Singh: A simple method of calculating individual subjective responses by paired comparisons tests. Memoirs of Graduate School of Science and Technology, Kobe University 14A, 57–66, 1996.
[40] M. Kendall: Further contributions to the theory of paired comparisons. Biometrics **11** (1955) 43–62.
[41] W. Clatworthy: Tables of two-associate-class partially balanced designs. Applied Mathematics Series/National Bureau of Standards, Washington, D.C., 1973.
[42] M. Glickman: Parameter estimation in large dynamic paired comparison experiments. Applied Statistics **48** (1999) 377–394.
[43] H. Scheffé: An analysis of variance for paired comparisons. Journal of the American Statistical Association **47** (1952) 381–400.
[44] R. Davidson, P. Farquhar: A bibliography on the method of paired comparisons. Biometrics **32** (1976) 241–252.

B

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 90 (2004) 161 – 170

Prediction of Subjective Response from Objective Measurements Applied to Walking Sound

Ann-Charlotte Johansson, Per Hammer, Erling Nilsson
Engineering Acoustics, LTH, Lund University, Box 118, 221 00 Lund, Sweden.
Ann-Charlotte.Johansson@acoustics.lth.se

Summary
This paper discusses prediction of the subjective response to walking sound – also called drum sound – based on differences in objective measurements. 'Walking sound' refers to the sound heard when someone is walking in the same room as the listener. Walking sound has attracted interest in recent years, particularly due to an increased use of thin floating floor constructions, such as veneer or laminate flooring, which can produce loud and sharp walking sound. A paired comparison test was performed in laboratory where listeners were asked which of the walking sounds was most disturbing. The response was analysed using a modified Bradley and Terry model allowing ties. Various measures, such as loudness according to ISO 532B, were tested against the subjective response using linear regression. The difference in 10-percentile loudness, N_{10}, between two stimuli was shown to predict the subjective response better than, for example, A-weighted sound pressure level. A difference of about 8% in N_{10} resulted in 50% of the subjects noticing a difference. The methodology used is applicable in situations when objective measures that have subjective counterparts are sought. Although the method is based on relative observations, an absolute ranking can be obtained by using a reference or a well-defined recording situation.
PACS no. 43.55.Hy, 43.66.Lj

1. Introduction

In this article subjective and objective measurements pertaining to walking sound are correlated to enable predictions of subjective opinions based on objective data. 'Walking sound' – sometimes also called *drum sound* – refers to the sound heard when someone is walking in the same room as the listener; hence not impact sound, although the correlation method might be applied to that type of sound as well. The objective is to find objective (measurable) parameters of the sound, such as loudness according to ISO 532B, that correlate well to the subjective impression of the disturbance caused by the walking sound. The goal is also to gain knowledge of the influence that differences in these objective measures has; that is, what quantity in the objective measure is needed to achieve a difference in the quality judgement.

Improving a product's sound at one time meant lowering the sound pressure of the radiated sound. The focus is finally changing, and now the emphasis is on improving the product's sound quality, which involves much more than simply lowering the sound pressure level. Product sound quality was defined by Jekosch and Blauert [1] as: 'Product sound quality is a descriptor of the adequacy of the sound attached to a product.' A motorbike should not sound like a vacuum cleaner; hence, it is not only by changing the perceived loudness that the product sound is improved. Other descriptors of the sound are needed. The field of psychoacoustics investigates the physical stimuli's effects on our hearing system. Parameters such as loudness, sharpness, fluctuation strength and roughness [2] are used to predict the subjective response.

In 1933, Fletcher and Munson presented a method of determining the loudness of complex tones [3]. By connecting points of perceived equal loudness level on a frequency-versus-sound-pressure-level diagram, equal-loudness or phon contours were introduced. In these contours, the frequency dependence of the ear's perception of a sound's strength is shown. One of several methods that appeared to include this effect in a measure of the perceived loudness, was the A-weighted sound-pressure level [dB(A)] which was derived from the 40-phon contour. In the same manner, dB(B) and dB(C) were introduced and derived from the 60- and 80-phon contours. As a result, these measures are applicable to certain levels of the sound. Predictions of the perceived loudness of sounds with various frequency and level contents based on, for example, the dB(A) measure are likely to fail.

The great diversity of procedures used to express the strength of sound forced ISO to harmonise the methods. As a first step, the dB(A) measure was standard-

Received 10 April 2003,
accepted 11 September 2003.

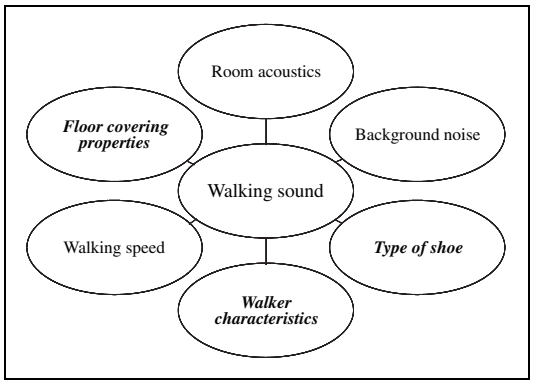

Figure 1. A complete model of how drum sound is perceived requires knowledge of several parameters. The italic parameters are investigated in the article.

ised, even though its limited applicability was well known. The loudness sensation, however, depends not only on the frequency content, other effects, such as masking and sound duration, have to be considered as well. More accurate, although more complex, predictions of the perceived strength of a sound were introduced a couple of years later in ISO 532 [4], where the measure of the perceived strength of sound is called loudness, and its unit is the sone. For the loudness measure, 1 sone is defined as the level of 40 dB of a 1-kHz tone. Two methods of calculating the measure loudness are described in ISO 532. In part A, loudness is calculated from octave-band analysis, while method B is calculated from one-third-octave band analysis. Part A, proposed by Stevens [5, 6, 7], and part B, proposed by Zwicker [8, 9], do not always agree. Part B generates generally higher results, but as it is said in the standard, method B seems to take better account of variations in narrow ranges of frequency of the sound spectra. The German standard, DIN 45 631 [10], corresponds to method B. Loudness in this case not only takes account of the level and frequency-dependency of the ear, but of the effects of masking and spectral distribution as well. The methods described in ISO 532 are meant for steady sounds and do not produce a time-variable measure, although temporal effects on loudness have been investigated and included in the loudness level meter by Zwicker [11]. Loudness is, unlike in the case of dB(A), a linear measure, i.e., a doubling of the loudness value produces a doubling of the sensation of loudness. This linearity is advantageous specifically in communicating differences and improvements to non-acousticians, and some examples are given in [12, 13]. Although the loudness measure in many cases has shown better correlation to subjective sensations of the sound strength than has dB(A), it is not as well known or used. This might be because dB(A)-weighting was introduced in a sound-level meter in 1936 [14], while the loudness measure was first introduced in a portable level meter in 1981 [15], although it was proposed in 1951 [16]. The time-consuming calculations were a problem in the past, but modern computers have reduced the effort needed. Increased awareness of the measure among acousticians seems to be needed in order to expand its use further.

In building acoustics, when evaluating the insulation of wall and floor structures it is mainly reference curves that are used to adjust physical measurements to performance in subjective listening tests [17, 18]. However, already in 1965 Fasold used loudness level when deriving an optimum theoretical curve for the normal impact sound level [19]. Psychoacoustic measures were also used in [20] to obtain quantitative figures on transmission loss for a facade. Good correlation to subjective preferences with regard to impact sound insulation was found in [21], by using a tapping machine and the ISO 532B loudness measure. In [22] the model of the human auditory-brain system proposed by Ando [23] was used to determine the correlation to the subjective evaluation, and in [24] various Zwicker parameters were included. It was found that a combination of loudness, sharpness, fluctuation strength, tonality and unbiased annoyance [25] correlated well to the subjective response to impact noise using the tapping machine.

In office environments several investigations have tested various measures pertaining to human perception of the sound from air-conditioning, office machines and human activities. In [26, 27] various measures, including the Zwicker and Stevens loudness level, were tested in an office environment; it was found that $L_{A,eq,T}$, where $T = 5$ min, was the best descriptor. Most investigations in office landscapes, however, have concerned speech.

Walking sound is not a new topic, it is mentioned in 1958 in [28], but it is only in the last 10 years that much interest in the topic has arisen. Thin floating floor coverings of, for example, laminate or veneer, have become common and can produce loud and sharp walking sound. The development of new floor coverings that successfully improves walking sound requires a thorough knowledge of how walking sound is perceived and how several parameters affect walking sound. The most important parameters appear in Figure 1. The floor coverings were assessed in the context of an office environment, where correlation between the surface and the sound is assumed to be less important than for domestic floors. It has been noticed by the author, that subjects, when knowing the design of the surface and when asked to choose a solution to their home, prefer different sound depending on the design of the surface. In an office environment the focus is however set to decrease the disturbance produced by walking sound. Thus, subjects were requested to imagine themselves in an office space and, in a paired comparison listening test, to say which of the sounds was the most disturbing.

In the initial tests of walking sound, semantic scales and paired comparison tests were used to find descriptors of the sound and to investigate the correlation of various objective measures to subjective preferences. The descriptor loudness correlated well to the perceived level of disturbance; the objective measures loudness (according to ISO

532B) and sharpness were seen valuable for predicting the level of disturbance [29].

For the subsequent tests (presented below), sounds were selected taking the results of the initial tests into account. Sound stimuli were chosen so that stimuli with a wide range of the psychoacoustic measures, loudness and sharpness, were represented. In a paired comparison test, 25 listeners indicated which of the sounds they perceived to be most disturbing. Ties – that is rating two sounds as equally disturbing – are permitted. Using a linear regression model, objective measurable data correlating to the results of the listening test were sought. Percentile values of loudness were calculated where the 10-percentile loudness showed very good correlation to the perceived disturbance, $R^2_{N_{10}} = 99.7\%$. Other objective measures such as sharpness did not in general improve the result.

2. Theory

2.1. Paired comparison test allowing ties

It is sometimes difficult for a subject to select an appropriate value from a descriptive scale, and for an experimenter to analyse the results obtained. There are uncertainties as to whether all subjects have equally understood and applied the scale. The problem is avoided using paired comparison tests where subjects are asked to judge which of two treatments has a certain attribute (e.g., a pleasant sound). There are several methods for analysing the results obtained from paired comparison tests, some of which are reviewed in [30]. Gridgeman recommends the following concerning whether or not to allow subjects to declare a tie [31]: When discrimination is the objective it is better to prohibit ties as the subjects' efficiency of decision might be offset, but when preference is the objective, ties should be allowed as they add information. Rao and Kupper [32] made modifications so as to allow ties to a paired comparison method formulated by Bradley and Terry [33, 34].

The Rao and Kupper model is based on the premise that when the difference between two treatments is smaller than a certain value, or threshold, the subjects will declare a tie. The probability of choosing T_i when compared to T_j is set to

$$P(T_i \to T_j) = \pi_{ij} = \int_{-(\ln \pi_i - \ln \pi_j)+\eta}^{\infty} \operatorname{sech}^2(y/2)\mathrm{d}y = \frac{\pi_i}{\pi_i + \theta\pi_j}. \quad (1)$$

π_i, $i = 1...t$, represent probability values for the t treatments, and η=ln(θ) is the sensory threshold for the subject. The probabilities for selection of j or for a tie when i and j are compared are calculated using the integral and are given in [32]. Figure 2 presents the expected outcome of the listening test. The influence of the sensory threshold value is shown in the figure as well.

These probabilities are here used in a linear regression to find psychoacoustic measures that correlate to the subjective judgements. It is assumed that the lower limit in the integral can be written as a linear measure based on differences in n objective measures, $\Delta X_{n,ij} = X_{n,i} - X_{n,j}$, between treatment i and j. That is

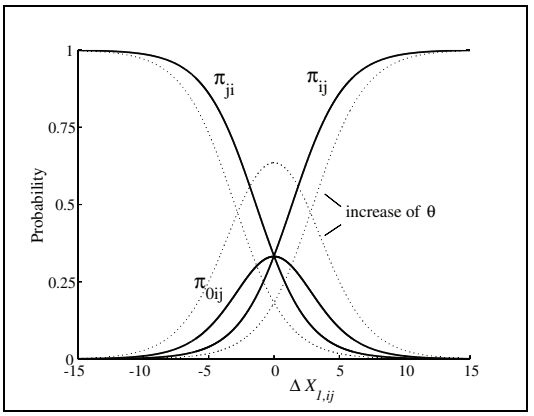

Figure 2. The expected outcome of the listening test when object *i* is compared with object *j*. π_{ij}, π_{ji} and π_{0ij} (equal disturbance) are plotted against the difference between the objects in a 'true' objective measure, $\Delta X_{1,ij}$. The dotted lines shows the influence of increasing the threshold value, θ.

$$\pi_{ij} = \int_{-f(\Delta X_{1,ij}, \Delta X_{2,ij}, \ldots, \Delta X_{n,ij})}^{\infty} \operatorname{sech}^2(y/2)\,\mathrm{d}y,$$

where

$$f(\Delta X_{1,ij}, \Delta X_{2,ij}, \ldots, \Delta X_{n,ij}) = \beta_0 + \beta_1 \Delta X_{1,ij} + \beta_2 \Delta X_{2,ij} + \ldots + \beta_n \Delta X_{n,ij}$$

and β_0, β_1 and β_2 are the linear regression coefficients.

The equation can be rewritten as

$$\pi_{ij} = \frac{1}{1 + e^{-f(\Delta X_{1,ij}, \Delta X_{2,ij}, \ldots, \Delta X_{n,ij})}} \quad (2)$$

or

$$-\ln\left(\frac{1}{\pi_{ij}} - 1\right) = \beta_0 + \beta_1 \Delta X_{1,ij} + \beta_2 \Delta X_{2,ij} + \ldots + \beta_n \Delta X_{n,ij}.$$

A disadvantage of the method is the rapidly increasing number of comparisons needed as the number of samples increases. The number of comparisons can be reduced using nearly balanced incomplete design techniques (balanced if the number of comparisons in which each sound is included is equal for all sounds, incomplete if all possible combinations are not represented).

3. Method

3.1. Sound stimuli

The choice of sound samples to be included in the test is critical for the results. If the test concerns sounds that have not been analysed before, it is particularly important that as many variations of the sound are represented as possible if general conclusions are to be drawn. The number of

Table I. Included samples and corresponding walker and shoes.

Drum sound	Floor	Walker shoe
1	14 mm parquet + PE foam	b
2	7 mm laminate + PE foam	b
3	7 mm laminate + fibreboard	b
4	10 mm laminate + fibreboard	b
5	10 mm laminate + glued PU underlay	a
6	7 mm laminate + underlay	a
7	10 mm laminate + glued PU underlay	d
8	22 mm pine wood + PE foam	c
9	Concrete	a
10	14 mm parquet + PU foam	c

a = female, hard high-heeled shoes, b = female, hard-heeled shoes, c = male, hard-heeled boots, d = male, hard-heeled shoes

samples is, however, often limited by practical considerations; the number of tests may become too large, and the desired sounds may not always be available.

In the initial test the floors tested represented common floor coverings in Sweden, such as linoleum, clinkers, parquet and laminate. The results of this test [29, 35] indicated that loudness according to ISO532B [4] and sharpness [36] were important tools in predicting the subjective opinion of the sound made by actual walkers. To make more accurate predictions, this investigation then focussed on using a wider set of drum sounds regarding loudness and sharpness. It is, however, impossible to keep all parameters but the floor coverings constant and still produce a wide range of drum sounds. Therefore, two different walkers were used, a female and a male, each wearing two different pairs of shoes. Recordings and measurements were made of several floor coverings, and combinations were selected that produced varying loudness, sharpness and frequency content. These are listed in Table I.

The influence of using various shoes and walkers was also investigated. In the case of varying the shoes, a female walker weighing 72 kg wore five different pairs of shoes each having different heel characteristics. In the case of varying the walkers, two female walkers weighing 64 and 70 kg and three male walkers weighing 80 to 95 kg were used. The female walkers wore the same women's shoes while the male walkers wore the same men's shoes, thus it was not only the walker that was changed. For each combination of shoes and walker, the walking sound on three of the floors were recorded. The floors used were no. 1, 3 and 5 in Table I. In the paired comparison test, subjects compared the three sounds made by each person or shoe; hence 30 comparisons were made (3 comparisons × 5 different shoes or walkers × 2).

3.2. Recordings

The sounds were recorded in a room of 162 m^3 having a reverberation time of 0.3 s at 200-315 Hz and 0.2 s at 400-5000 Hz. Background noise consisted of slight fan noise, less than 35 dB (one-third-octave band level) up to 500 Hz, thereafter less than 25 dB. The samples listed in Table I were laid on the bare concrete floor in the middle of the room, see Figure 3. The walker walked back and forth over the sample at a speed of 2 steps/s for 30 s. Two-channel recordings of the walking sounds were made using a dummy head and B&K 4181 microphones, G.R.A.S. 26AK preamplifiers and the digital real-time analyser 01-dB Stell Symphonie (sampling frequency 51.2 kHz). The ears' height above the floor surface was 1.15 m – as for a sitting person.

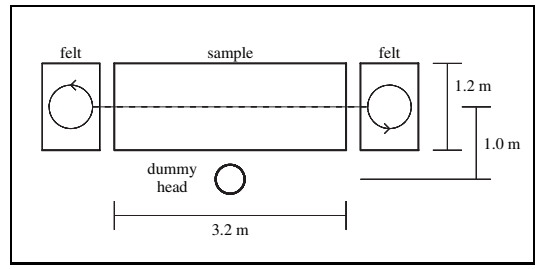

Figure 3. Measurement set-up.

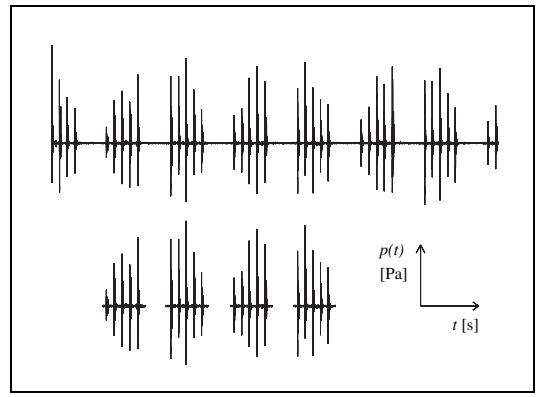

Figure 4. From the 30 s recordings four samples of 2.8 s duration were selected for inclusion in the test.

3.3. Treatment of recordings

From the 30 s recordings four samples were selected for inclusion in the test, as in Figure 4. The duration of each selection was 2.8 s. The four first 'five step sections' available were selected unless irregularities in the walking, such as a clear stumble, were perceived in any of them.

3.4. Playback

During the listening tests the subjects listened to the recordings using headphones. It was found in earlier tests [29], by direct comparison of live and recorded drum sound, that supra-aural headphones gave the most natural reproduction. The STAX SR-80 Pro headphones, the NAD 312 amplifier and MATLAB were used. The playback level using headphones was checked by direct comparison of live and recorded drum sound. The calibration

was performed by seven subjects: the agreement of the seven was good, and the average value of the chosen playback level was used.

3.5. Paired comparison test

The listening test was performed in a conference room of 5.8 × 6.0 × 2.8 m^3. The background noise from the computer was minimised. The subjects were introduced to the test and familiarised with the sounds before the test started. A mock-up example without any collection of results was conducted as well. Subjects were thereafter requested to imagine themselves in an office space, and in a paired comparison test, tell which of the sounds was most disturbing (in Swedish '*störande*'), see Figure 5. It was possible to declare a tie. The subjects could switch between the sounds as many times as they wanted to, with no time limit. As a button was pushed, a sound sample of five steps sounded. The sound sample was chosen randomly from among the available five-steps intervals from the original 30 s recordings, see Figure 4.

The test followed a balanced incomplete design with seven replicates of the 10 original sound samples, resulting in 35 comparisons. The 30 comparisons of varying shoes and walkers were tested at the same time, so 65 comparisons were made.

The order of the sound samples was randomised for each subject. Each subject compared each pair of sound samples once, divided between two sessions of approximately 20 min each. During the test, the test subjects were left alone, although they could get the attention of the test leader if necessary. The test results were saved automatically in data files for further analysis.

Thirteen female and 12 male subjects were used: 16 were members of a choir at Lund University and the remainder worked in the department. Their ages ranged from 21 to 60, although over 80% were between 21 and 35 years old. None of the subjects reported any hearing disabilities.

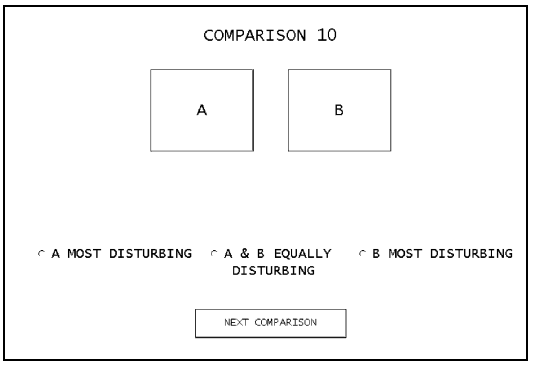

Figure 5. The paired comparison test was performed using MATLAB. As a button was pushed, a sound example of five steps was played. The test results were saved automatically. The figure is a translation of the Swedish version.

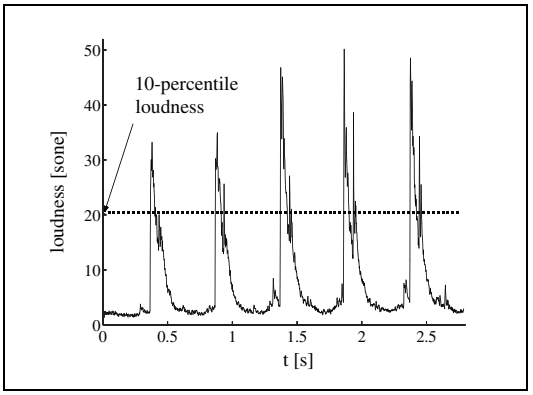

Figure 6. Loudness as a function of time calculated with an integration time of 1 ms. The figure also shows the 10 percentile, that is the value that is exceeded 10% of the time.

3.6. Objective psychoacoustic measures

The 01-dB Stell Symphonie was used to calculate loudness, N, according to ISO 532B based on equivalent one-third-octave band levels, L_{eq}, for the entire signal (T=30 s) as it was done in earlier investigations [29]. Every calculation was made on the each channel recording separately. The mean of the two channels' result was taken to be representative.

$$L_{eq,T} = 10 \log \left(\frac{1}{T} \int_0^T 10^{L_p(t)/10} \mathrm{d}t \right) .$$

As only some parts of the whole signal was used, the same calculations for the four selected portions of each signal (each of 2.8 s duration) were made where the median value was taken to be representative. The equivalent A-weighted and C-weighted sound pressure levels, L_A [dB(A)] and L_C [dB(C)], were also calculated for the selected parts with T=2.8 s. No corrections of the measured loudness due to binaural hearing as in [37] were made.

Loudness percentiles (1 to 99 percentile), sharpness and sharpness percentiles were calculated using 01-dB Stell Symphonie and MATLAB, Figure 6. Sharpness, S, was calculated according to von Bismarck [38] and Zwicker [36] based on the received loudness pattern using the entire signal (T=30 s). The percentiles were calculated for the four selected portions of each signal according to section 3.3 and the median value was taken to be representative. Loudness as a function of time was calculated for using an integration time step of 1 ms, where loudness was given from the specific loudness pattern of each step. The 1 ms time step was the smallest possible integration time supported by the software, and it was chosen so as to reduce the smoothing effect of the integration. No effects of pre- and post-masking are included and the temporal envelope of the basilar membrane is not represented here. Sharpness as a function of time was calculated using the

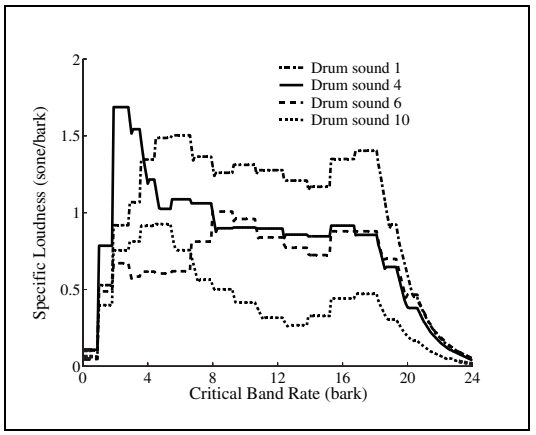

Figure 7. Loudness patterns for some drum sounds. Loudness calculations were made using one-third-octave band levels, L_{eq}, for the entire signal (30 s).

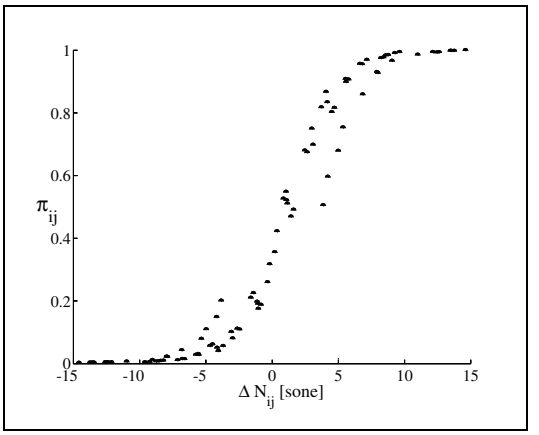

Figure 8. The probability that object *i* is perceived to be more disturbing than object *j*, π_{ij}, is plotted against the difference in loudness, ΔN_{ij}.

loudness patterns that were calculated to receive loudness as a function of time.

Every calculation was made on each of the two channel recordings.

4. Results

4.1. Subjective listening test and objective measurements

Loudness using the entire signal, N, 10-percentile loudness, N_{10}, sharpness, S, 90-percentile sharpness, S_{90}, L_A, L_C and probability values, π_i, from the initial test are listed in Table II. Loudness calculations on the selected signals of duration 2.8 s produced similar results as when using the entire signal, N, and are therefore excluded. Loudness patterns for four of the drum sounds are shown in Figure 7 to show some examples of the sounds' character.

The standard deviation for the four selected portions of each signal regarding the 10-percentile loudness was between 0.4 and 1.7 sone, in average 4% of the sone value.

The results are shown in Figure 8-11 where the probability that object *i* is perceived to be more disturbing than object *j*, π_{ij}, is plotted against the difference in objective measures N, N_{10}, L_A and L_C.

The statistic R^2 [39] is a measure of the reduction in variability of the outcome when using the regression variables. When several variables are included, the adjusted R^2 is helpful in ensuring that an increase of R^2 is not due to the inclusion of extra factors in the regression model. The percentile of loudness that correlated the best to the subjective opinion regarding R^2 was shown to be the 9 percentile. R^2 for various combinations of objective measures are listed in Table III, as are the regression coefficients. The adjusted R^2 in the case of including $\Delta N_{10,ij}$ and ΔS_{ij} equals 99.7 %.

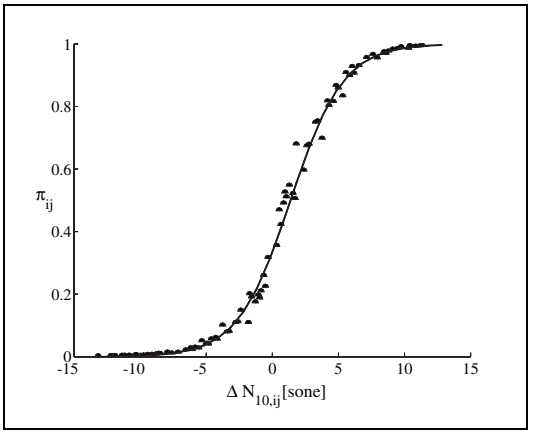

Figure 9. The probability that object *i* is perceived to be more disturbing than object *j*, π_{ij}, is plotted against the difference in 10 percentile loudness, $\Delta N_{10,ij}$. The solid line was calculated according to equation 3.

The regression line of including only $\Delta N_{10,ij}$ in equation (2) is included in Figure 9 and can be written

$$\pi_{ij} = \frac{1}{1 + \mathrm{e}^{-(-0.69+0.50\Delta N_{10,ij})}}. \qquad (3)$$

Corresponding lines for π_{ji} and π_{0ij} (equal disturbance) are shown by the solid lines in Figure 2.

The results of the test varying the walker and the shoes are listed in Table IV. The shoes are listed from the lowest heel to the highest heel (Sh1-Sh5), while the walkers are listed from the lightest to the heaviest (W1-W5). Comparisons between the probability values, and loudness and sharpness values in the table should be made within a row. Thus for Sh1, floor covering 3 was perceived as producing the most disturbing sound and it had the highest N_{10}-value. Probability values were calculated according

Table II. Probability values, loudness N, 9-percentile loudness N_{10}, sharpness S, 90-percentile sharpness S_{90} and the A-weighted sound pressure level L_A. θ from the Rao-Kupper model was found to be 2.0.

Drum sound	Floor	Walker shoe	Probability values	N [sone]	N_{10} [sone]	S [acum]	S_{90} [acum]	L_A [dB(A)]	L_C [dB(C)]
1	14 mm parquet + PE foam	b	0.51	24.8	23.9	1.35	1.15	71.4	72.0
2	7 mm laminate + PE foam	b	0.25	23.2	22.8	1.48	1.29	70.7	70.9
3	7 mm laminate + fibreboard	b	0.12	19.8	21.1	1.29	1.11	67.0	69.8
4	10 mm laminate + fibreboard	b	0.08	19.5	20.5	1.23	1.00	65.8	69.8
5	10 mm laminate + glued PU underlay	a	0.02	16.8	17.8	1.28	1.15	65.6	67.3
6	7 mm laminate + underlay	a	0.01	15.8	16.3	1.48	1.33	65.2	65.0
7	10 mm laminate + glued PU underlay	d	0.002	12.7	12.5	1.42	1.32	61.3	62.2
8	22 mm pine wood + PE foam	c	0.001	11.3	12.0	1.02	0.91	60.2	65.4
9	Concrete	a	0.001	11.1	11.7	1.36	1.08	57.3	59.5
10	14 mm parquet + PU foam	c	0.001	10.2	10.7	1.15	1.02	57.8	61.3

a = female, hard high-heeled shoes, b = female, hard-heeled shoes, c = male, hard-heeled boots, d = male, hard-heeled shoes

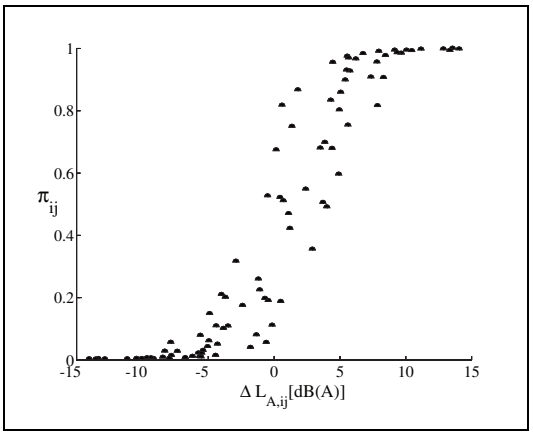

Figure 10. The probability that object *i* is perceived to be more disturbing than object *j*, π_{ij}, is plotted against the difference in A-weighted sound pressure level, $\Delta L_{A,ij}$.

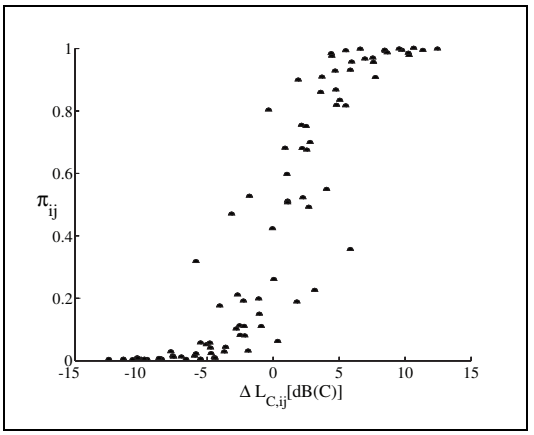

Figure 11. The probability that object *i* is perceived to be more disturbing than object *j*, π_{ij}, is plotted against the difference in C-weighted sound pressure level, $\Delta L_{C,ij}$.

Table III. R^2 and regression coefficients for various linear combinations of objective measures; adjusted R^2 when including both the differences in N_{10} and S equals 99.7 %.

$f(\Delta X_{1,ij}, \Delta X_{2,ij})$	R^2 [%]	β_0	β_1	β_2
$\beta_0 + \beta_1 \Delta N_{10,ij} + \beta_2 \Delta S_{ij}$	99.7	-0.69	0.50	0.27
$\beta_0 + \beta_1 \Delta N_{10,ij}$	99.7	-0.69	0.50	
$\beta_0 + \beta_1 \Delta N_{ij}$	98.1	-0.69	0.47	
$\beta_0 + \beta_1 \Delta L_{A,ij}$	93.4	-0.69	0.49	
$\beta_0 + \beta_1 \Delta L_{C,ij}$	91.2	-0.69	0.55	

to Rao and Kupper, but with a suppression of θ equal to the threshold value from the original test. Since θ can be seen as the sensory threshold for the subjects and as the tests were performed at the same time, θ should stay equal, see section 5. Figure 12 shows the positions of the different variations using the difference in N_{10}.

5. Discussion

L_A and L_C did not correlate as well to the results of the listening test as did N and N_{10}. Even though much of the variability of the results can be explained by L_A, the variability is better explained by N and N_{10}. In other words, even though loudness according to ISO 532B, herein called N, is meant to be used for stationary sounds, it is still better than L_A when applied to drum sound.

The 10-percentile loudness, N_{10}, was shown to correlate the best of all presented measures to the subjective opinion, and it can apparently predict the subjective response well. The inclusion of sharpness in the regression model did improve the statistic R^2, but by less than 0.1 percentage point. Furthermore, the contribution of S was small, as N_{10} and S differ in most comparisons by more than a factor of 10 in the objective measurements; its contribution is also small in light of the measurement uncertainties. When the difference in N_{10} is small, S might have

Table IV. Probability values and 10-percentile loudness for the test varying shoes (Sh1-Sh5) and walkers (W1-W5). Comparisons between probability values and 10-percentile loudness should be made within a row. Floor 1 = 14 mm parquet + PE foam, 3 = 7 mm laminate + fibreboard, 5 = 10 mm laminate + glued PU underlay.

	Probability values, π_i			N_{10} [sone]		
Floor	1	3	5	1	3	5
Sh1	0.17	0.73	0.09	13.4	16.1	13.0
Sh2	0.39	0.31	0.30	20.1	20.0	18.6
Sh3	0.43	0.51	0.06	17.7	18.3	13.3
Sh4	0.33	0.60	0.07	22.7	24.4	18.7
Sh5	0.26	0.59	0.15	21.5	24.9	19.0
W1	0.46	0.33	0.21	21.5	21.2	18.9
W2	0.73	0.19	0.08	26.6	25.7	22.5
W3	0.50	0.33	0.17	10.6	9.9	8.4
W4	0.53	0.31	0.16	15.7	15.5	12.8
W5	0.27	0.60	0.14	10.2	11.9	8.9

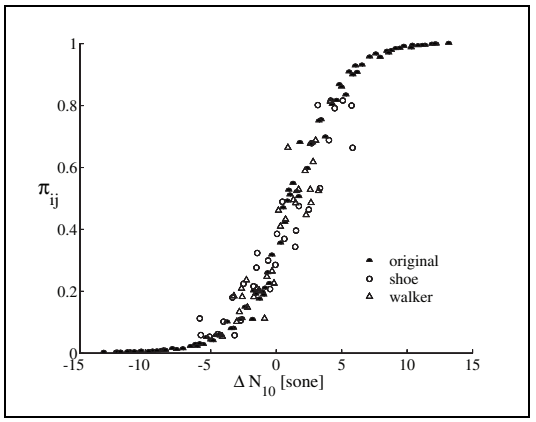

Figure 12. The probability that object *i* is perceived to be more disturbing than object *j*, π_{ij}, is plotted against the difference in 10 percentile loudness, $N_{10,ij}$. The circles and triangles are, respectively, the results obtained by varying shoes and walkers.

a larger influence but the sound samples are here too few to draw any conclusions.

S, i.e. the sharpness calculated from the loudness pattern of the entire signal using L_{eq} with T=30 s, was chosen as the second variable. There was no difference in R^2 when using sharpness percentiles (1 to 99 percentiles) instead of S. (The profile of sharpness versus time is reversed compared with that of loudness in Figure 6; i.e., during the 'silent' periods, sharpness is higher than during the impacts. Therefore the 90 percentile, i.e., during 10% of the time the sharpness was lower than S_{90}, was chosen to be presented in Table II.)

Of all percentile values, the 9-percentile loudness resulted in the highest R^2-value ($R^2_{N_9} = 99.73\%$ compared with $R^2_{N_{10}} = 99.68\%$). N_{10} was nevertheless chosen as it is used in the annoyance measure of Zwicker [25]. Other loudness percentiles have, however, been introduced as well. In the measure of psychoacoustic annoyance proposed by Widmann, the 5-percentile is included [40]. Fastl found that the 4-percentile loudness corresponded to the perceived average loudness of road traffic noise [41]. Both $R^2_{N_5}$ and $R^2_{N_4}$ were found to equal 98.0%, hence N_{10} and N_9 are better descriptors in this case. (Note that $R^2_{N_4}$ in [41] was calculated differently than it was in this paper.) In the loudness meter proposed by Zwicker [11], the effects of pre- and post-masking are included and the temporal envelope of the basilar membrane is represented by using a RC network with a time constant of 2 ms. A simpler treatment which excludes these effects was used here, but the results nonetheless correlate well to the subjective response. A better model of the ear would be unlikely to yield better predictions due to the uncertainties in the measurement set-up.

The correlation between objective measures and subjective response was lower when using various shoes and walkers. It might indicate that other unknown measures not tested here are needed. Figure 12 shows the result obtained by suppressing θ so as to equal the threshold value from the original test; without suppression, the result would appear as it does in Figure 13. Suppression is justified as follows. The threshold value is associated with the sensory threshold that needs to be exceeded in order for the subject to notice a difference. This threshold should remain constant during the test. In this test various differences were used and they were randomly presented. (Using only small differences throughout a test could result in subjects sharpening their hearing acuity.) The different sounds obtained by varying the shoes and walker were, however, sometimes very similar. It is possible that the subjects hesitated about their choice and sometimes made the 'wrong' decision, which resulted in a higher threshold value in the Rao-Kupper model. If more subjects are used, the variability decreases and the threshold value decreases. It was judged likely that the threshold value should decrease to equal that of the original test if more subjects were used and therefore θ was suppressed.

For a comparison with earlier results found in [29], π_{ij} was calculated as well as the difference in N. The spread of the results proved to be much wider when applying the coefficients β_0 and β_1 found in here. In [29] a different question was posed: 'Which sound is most pleasant?' The recording situation was less controlled and there were possibly too few listeners (subjects). It can, however, indicate that more measures are needed, especially when the difference in N is small, even though they seem unnecessary in the latest findings. N_{10} would probably improve the results, but unfortunately no calculations on the difference N_{10} could be made due to the different measurement technique used.

When planning listening tests, a question that often arises is how many listeners are needed? Figure 14 shows the probability values π_i for the 10 drum sounds in percent of the final value as a function of the number of listeners. The variation has decreased, but some variation is still ev-

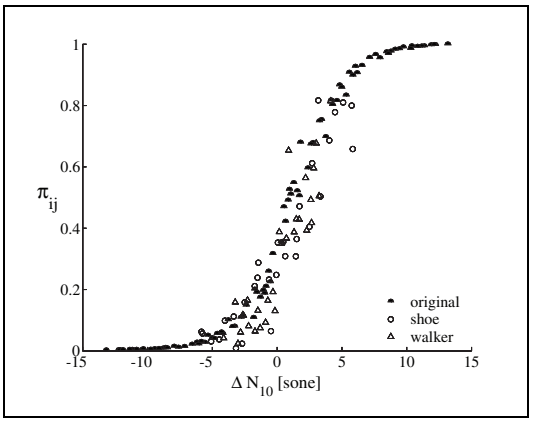

Figure 13. No suppression of θ. The probability that object *i* is perceived to be more disturbing than object *j*, π_{ij}, is plotted against the difference in 10 percentile loudness $N_{10,ij}$. The circles and triangles are, respectively, the results obtained by varying shoes and walkers.

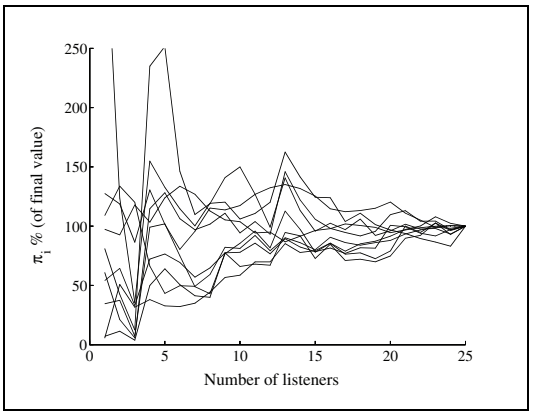

Figure 14. The probability values π_i for the 10 drum sounds in percent of the final value as a function of the number of listeners.

ident even when the number exceeds 22 listeners. It would have been safer to use 30 listeners to see how the variation stabilises, especially for the cases of varying the shoes and walker. The results shown in Figure 9 and 13 would then perhaps be better.

The result can be used to estimate how large difference in N_{10} is needed to produce a 50% probability that one of the paired walking sounds is perceived to be more disturbing than the other. Note that ties are permitted, hence the 50% probability is not at a difference of 0, see Figure 2. Using equation (3) or Figure 9 indicates that, in this case, a difference in N_{10} of 1.4 sones results in the probability π_{ij} equalling 50%. The 10-percentile loudness values were between 10.7 and 23.9 sone with a median value of 17.1 sone, 1.4 sone corresponds therefore in average to a difference of 8%. Terhardt found in [42] when investigating the just noticeable difference value using pulses of a 1-kHz tone with levels of 40, 60 and 80 dB, that a loudness difference of 9% is required. Even though the methods and sounds that are used in these investigations differ, the results are surprisingly similar. It would be interesting to know if the result is the same for other types of sound, and thus can be seen as a general rule.

The prediction presented here is based on relative observations in a controlled environment. A control of the result's applicability in field is desirable but has not been performed. In a field study it might be necessary to include the effects of different room acoustics and background noise in the model. The field study could contain semantic scales so that the degree of disturbance could be investigated, from a paired comparison test the ranking of the sounds is known but not the degree of the disturbance.

Based on objective measurements prediction of the subjective response when compared to another sound can be made using equation (3). An absolute measure can be obtained by using a well-defined reference or recording situation. Though this is not the goal of this study, it might be of interest in other applications. In cases where a clear reference is sought and available, a range of classes of walking sounds can be created by means of defining various probability percentages, such as 50, 70 and 90%, so that each class represent a certain amount of people evaluating a difference to the reference. The extent of a class can also be based on a specific probability percentage. As an example, if the percentage of 80 is chosen for walking sound, it would mean that each class would comprise 4 sone according to equation (3). Hence, in both cases different classes would then not only correspond to different numeric values but also have a subjective counterpart.

6. Concluding remarks

It was shown that a difference in 10-percentile loudness N_{10} can predict the subjective disturbance when two walking sounds are compared in a laboratory environment. Sharpness does not improve the prediction to any significant extent. N_{10} is a much better objective measure than L_A and L_C in particular, and better than loudness calculated at one-third-octave band levels, L_{eq}, for the entire signal (30 s). The result, that percentile values are better than representing the sound based on the arithmetic mean value of the signal's sound energy, agrees with earlier findings regarding impulse sounds [41].

A difference of about 8% in N_{10} resulted in 50% of the subjects noticing a difference. This result support findings by Terhardt [42].

The methodology used in this article is naturally applicable in other situations when objective measures that have subjective counterparts are to be found. Although the method is based on relative observations, an absolute ranking can be obtained by using a reference or a well-defined recording situation.

Acknowledgement

The authors wish to acknowledge the help and support of their colleagues, especially Dr. Jonas Brunskog, at the di-

vision of Engineering Acoustics, LTH, Lund University, Sweden. The authors also wish to thank the Swedish Foundation for Knowledge and Competence Development and Pergo AB for their financial support.

References

[1] U. Jekosch, J. Blauert: A semiotic approach toward product sound quality. Inter-Noise 96, GB-Liverpool, 1996, 2283–2286.
[2] H. Fastl: The psychoacoustics of sound-quality evaluation. Acustica/acta acustica **83** (1997) 754–764.
[3] H. Fletcher, W. A. Munson: Loudness, its definition, calculation and measurement. J. Acoust. Soc. Am. **5** (1933) 82–108.
[4] ISO 532: Method for calculating loudness level. 1975.
[5] S. Stevens: The measurement of loudness. J. Acous. Soc. Amer. **27** (1955) 815–829.
[6] S. Stevens: Calculation of loudness of complex noise. J. Acous. Soc. Amer. **28** (1956) 807–832.
[7] S. Stevens: Procedure for calculating loudness: Mark VI. J. Acous. Soc. Amer. **33** (1961) 1577–1585.
[8] E. Zwicker: Subdivision of the audible frequency range into critical bands (frequenzgruppen). J. Acous. Soc. Amer. **33** (1961) 248.
[9] E. Zwicker: A model of loudness summation. Psychological Review **72** (1965) 3–26.
[10] DIN 45631: Procedure for calculating loundness level and loudness. 1991.
[11] E. Zwicker: Procedure for calculationg loudness of temporally variable sounds. J. Acous. Soc. Amer. **62** (1977) 675–682.
[12] E. Zwicker: What is a meaningful value for quantifying noise reduction? Inter-Noise 85, Munich, 1985, 47–56.
[13] E. Zwicker: Meaningful noise measurement and effective noise reduction. Noise Control Engineering Journal **29** (1987) 66–76.
[14] ASAZ24.3: American tentative standards for sound level meters for measurement of noise and other sounds. J. Acoust. Soc. Am. **8** (1936) 147–152.
[15] E. Zwicker, W. Daxer: A portable loudness meter based on psychoacoustical models. Inter-Noise 81, Amsterdam, 1981, 869–872.
[16] L. L. Beranek, J. L. Marshall, A. L. Cudworth, A. P. G. Peterson: Calculation and measurement of the loudness of sounds. J. Acous. Soc. Amer. **23** (1951) 261–269.
[17] ISO 717: Rating of sound insulation in buildings and of building elements. 1996.
[18] K. Bodlund: Alternative reference curves for evaluation of the impact sound insulation between dwellings. Journal of Sound and Vibration **102(3)** (1985) 384–402.
[19] W. Fasold: Untersuchungen über den Verlauf der Sollkurve für den Trittschallschutz im Wohnungsbau. ACUSTICA **15** (1965) 271–284.
[20] E. Zwicker, H. Fastl: Examples for the use of loudness: Transmission loss and addition of noise sources. Inter-Noise 86, Cambidge, Massachusetts, 1986, 861–866.
[21] E. Nilsson, P. Hammer: Subjective evaluation of impact sound transmission through floor structures. Tech. Rept. TVBA-3103, Engineering Acoustics, Lund University, 1999.
[22] J. Y. Jeon: Subjective evaluation of floor impact noise based on the model of ACF/IACF. Journal of Sound and Vibration **241** (2001) 147–155.
[23] Y. Ando: A theory of primary sensations and spatial sensations measuring environmental noise. Journal of Sound and Vibration **241** (2001) 3–18.
[24] J. Y. Jeon, J. H. Jeong, Y. Ando: Objective and subjective evaluation of floor impact noise. Journal of Temporal Design in Architecture and the Environment **2** (2002) 20–28.
[25] E. Zwicker: A proposal for defining and calculating the unbiased annoyance. Contributions to Psychological Acoustics, A. Schick et al. (eds.) Bibliotheks- und Informationssystem der Universit t Oldenburg, p. 187-202, 1991.
[26] S. K. Tang: Performance of noise indices in air-conditioned landscaped office buildings. J. Acous. Soc. Amer. **102** (1997) 1657–1663.
[27] U. Ayr, E. Cirillo, I. Fator, F. Martellotta: A new approach to assessing the performance of noise indices in buildings. Applied Acoustics **64** (2003) 129–145.
[28] O. Brandt: Akustisk planering, In Swedish. Stockholm, 1958.
[29] A.-C. Johansson, E. Nilsson, P. Hammer: Footstep sound from different floor coverings, subjective measurements. Acoustics 2000, Volume 22, GB-Liverpool, 2000, 95–100.
[30] A.-C. Johansson, P. Hammer, E. Nilsson: Aspects on three methods for paired comparison listening tests. ICA 01, Rome, 2001, 2 pages.
[31] N. Gridgeman: Pair comparison, with and without ties. Biometrics **15** (1959) 382–388.
[32] P. V. Rao, L. L. Kupper: Ties in paired comparison experiments: A generalization of the Bradley-Terry model. J. Amer. Stat. Assoc. **62** (1967) 194–204, Corrigenda 63 1550.
[33] R. A. Bradley, M. E. Terry: The rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika **39** (1952) 324–345.
[34] R. A. Bradley: Paired comparisons: Some basic procedures and examples. – In: Handbook of Statistics 4: Nonparametric Methods. P. R. Krishnaiah, P. K. Sen (eds.). Elsevier Science Publishers, 1984, 299–326.
[35] A.-C. Johansson, E. Nilsson, P. Hammer: Subjective classification of footstep sound from different floor coverings. Inter-Noise 2000, Volume 5, France, Nice, 2000, 3248–3251.
[36] E. Zwicker, H. Fastl: Psychoacoustics, facts and models, 2nd Ed. Springer-Verlag, Berlin, 1999.
[37] E. Zwicker, U. T. Zwicker: Dependence of binaural loudness summation on interaural level differences, spectral distribution, and temporal distribution. J. Acous. Soc. Amer. **89** (1991) 756–764.
[38] G. von Bismarck: Sharpness of steady sounds. ACUSTICA **30** (1974) 159–172.
[39] D. C. Montgomery: Design and analysis of experiments, 5th ed. John Wiley & Sons, New York, 2001.
[40] U. Widmann: Aurally adequate evaluation of sounds. Proc. Euro Noise 98, 1998, 29–46.
[41] H. Fastl: Evaluation and measurement of perceived average loudness. – In: Contributions to Psychological Acoustics. A. S. et al. (ed.). Bibliotheks- und Informationssystem der Universit t Oldenburg, 1991, 205–216.
[42] E. Terhardt: Über ein Äquivalenzgesetz für Intervalle akustischer Empfindungsgrößen. Kybernetik **5** (1968) 127–133.

C

Evaluation of Drum Sound with ISO Tapping Machine

Ann-Charlotte Johansson, Erling Nilsson, Per Hammer

Division of Engineering Acoustics, LTH, Lund University,
P.O Box 118, SE 221 00 Lund, Sweden
Ann-Charlotte.Johansson@acoustics.lth.se

ABSTRACT
A branch norm, EPLF NORM 021029-3 [1], has been established for measuring drum sound on laminate floor coverings. 'Drum sound' refers to the sound occurring when an object, e.g. a foot, strikes the flooring in the room in which the receiving ear is located. The norm evaluates the subjective perception of the drum sound's loudness using the ISO tapping machine. A round-robin study of the norm is reported along with the results of a paired comparison listening test using the same floor coverings. The article discusses general aspects of evaluation measures, tapping machines, test environments, etc., that need to be considered when measuring drum sound on various floor coverings, such as linoleum, wood parquet and laminate. It is concluded that loudness as measured according to ISO 532B correlates the best with the subjective perception of the drum sound's loudness. The tapping machine can be used to excite hard floor coverings to produce the drum sound, but should be used with caution in studying low-level drum sounds due to the tapping machine's inherent mechanical noise.

1 INTRODUCTION

Interest in the effect of background noise on health and work capacity has increased. One disturbing factor in office spaces, hotels, schools, etc., is the 'drum sound', i.e., the sound produced when an object, such as a foot, hits the flooring in the room in which the receiving ear is located. Drum sound is sometimes also called 'walking sound' or 'drum noise', although 'drum noise' should be avoided since the term 'noise' suggests that the sound is unwanted, which is not always the case. Drum sound has attracted interest in

recent years, particularly due to the increased use of thin floating floor constructions, such as veneer or laminate flooring, which can produce loud, sharp drum sounds when a person wearing hard-heeled shoes walks on them. Consumers' increasing demand for floorings with improved drum sound properties has made the industry interested in producing better products. The Division of Engineering Acoustics at Lund University in Sweden has worked in the field since the early 1990s. Recordings, listening tests and measurements of live walkers have been used to determine what causes disturbance, and to improve the drum sound. Johansson et al. [4] have developed a measure that correlates the subjective perception of the disturbance of the drum sound to recordings of the drum sound. The first study of drum sound known to the authors was made in Denmark in 1952 by Larris [5]. In that study, measurements were made of the sound produced by a tapping machine on various types of floorings, and listening tests were performed in which listeners judged the level of the sound of a person walking on the floor by the help of a Barkhausen phon-meter (generating an 800 Hz tone). Two alternatives are suggested for representing the drum sound as a single figure: the mean value in the 50 to 800 Hz interval, or the mean value in the frequency interval of maximum sound pressure level.

As better products are developed, a standard method for measuring floor performance and presenting it to the market is needed. Many measurements methods are used and there is a need to harmonise them. Naturally, the final method must be repeatable, reproducible, practical and must correspond to subjective perceptions of the sound – something surprisingly often neglected. In 2001 the European Producers of Laminate Floorings (EPLF) started to search for suitable methods. The Division of Engineering Acoustics at Lund University presented a method that was chosen for initial consideration [2]. This method uses the ISO tapping machine and measurements are made in damped rooms with the flooring of interest laid on a concrete floor. The measured loudness [3] is used in the evaluation, as earlier investigations found high correlation between loudness and the subjective perception of the drum sound. The norm has, however, been revised since then. In early 2004 a final round-robin test was performed. In conjunction with this, a paired comparison listening test was also performed in the laboratory at the Division of Engineering Acoustics. In this test listeners were asked which drum sound was the loudest; this was done to control for the correlation of the norm to subjective response. The results of the round-robin and listening test are reported here.

An ad-hoc group was established in 2003 following a CEN TC 126 resolution to prepare a standard for 'Laboratory measurement of walking noise on floors'. The intention is so far to keep the standardised tapping machine (with or without modifications) as the source of the drum sound, and to use a test room and measurement set-up similar to that

of the EN ISO 140 series. More information on other methods used in Nordic and other European countries can be found in [6]; two of these methods are presented below.

The French standard, NF S 31-074: 'Laboratory measurement of in-room impact noise by floor covering in this room' [7], approved in 2002, is based on EN ISO 140-8 [8]. Measurements are made in a reverberation room. To reduce the inherent mechanical noise, a sound insulating cover is placed over the tapping machine. The result is presented using the $L_{n,e,w}$ index, evaluated in accordance with EN ISO 717-2 [9]. The method is not intended to provide a subjective evaluation of the results.

DELTA Danish Electronics, Light and Acoustics employs a method similar to that of EN ISO 140-8. The test floor is laid in a reverberation room on the standard concrete floor. A standardised tapping machine is used. A reverberation time correction is applied, and the normalised drum sound level per 1/3-octave band is given as the result without any evaluation of a single numeric value [6].

This paper discusses general aspects of evaluation measures, tapping machines, test rooms, test samples, etc., that have to be dealt with when considering measurement of drum sound on various floor coverings, such as linoleum, wood parquet and laminate, and in seeking a suitable norm. After this, a short description of the branch norm, EPLF NORM 021029-3 [1], for laminate floor coverings and its scope is given along with the results of the round-robin test of the norm and the corresponding listening test.

2 GENERAL ASPECTS

2.1 Evaluation measure

The aim that an objective value should have a subjective counterpart could mean many things. Is it a measure of, for example, perceived loudness, disturbance, pleasantness and/or pitch that is desired? Should the measure indicate improved product sound quality? Product sound quality was defined by Jekosch and Blauert [10] as: 'a descriptor of the adequacy of the sound attached to a product'. Regarding flooring coverings, this could mean that the drum sound from a wood floor should not sound like that of a stone floor. Is it then even possible to formulate a common norm for all floor coverings? If the starting point is to improve, for example, the office environment by decreasing the disturbance caused by people walking on the floor, it can be assumed that it is not important to the listener whether the sound has a 'wooden' or 'stone' character; what is important is that the sound causes disturbance and that the disturbance should be decreased. In other cases, the character of the sound may be more important, and a

decrease of the amplitude of the drum sound should then be accomplished without negatively affecting the character of the sound. A study by Johansson and Nilsson [11] found perceptions of the disturbance and loudness of a drum sound to be highly correlated (correlation coefficient = 0.995) when listeners were judging recordings of drum sounds. Therefore, by choosing a measure that correlates to the perceived loudness it is possible to produce a true and clear measure for the office and home environments. If needed, such a measure can be complemented with another measure describing the character of the sound. However, a measure describing the pitch has yet to be found, 'sharpness' [12] having been tested without success.

The present authors investigated in [4] the correlation between the subjective perception of disturbance produced by the drum sound and recordings of the same drum sound. Listeners were asked in a paired comparison test performed in a laboratory which recorded drum sound was most disturbing. Various measures, such as loudness (sone) according to ISO 532B [3], were tested against the subjective response using linear regression. The difference in loudness between two stimuli was shown to predict the subjective response better than, for example, A-weighted sound pressure level did. The R^2 statistic, a measure of the reduction in variability of the outcome when using the regression variables, was 98.1% for loudness and 93.4% for A-weighted sound pressure level [4]. In another study the listeners again listened to recordings of people walking and gave their judgements, but the objective measures were obtained using the ISO tapping machine as the source of the drum sound. A larger range of flooring materials was included, so the differences between the studied drum sounds were clearer than in previous studies. In that case, C-weighted sound pressure level showed as good a correlation as loudness did to the perceived loudness as well as to the perceived disturbance [11]. In the round-robin study and associated listening test reported on below, loudness according to ISO 532B showed higher correlation to the perceived loudness than did either A- or C-weighted sound pressure level. The correlation coefficient in this case was 0.83 using the loudness measure, whereas it was 0.72 and 0.73, respectively, using the A- and C-weighted sound pressure levels. This indicates that when the level differences between drum sounds are great, a rough measure such as C-weighted sound pressure level can be just as good as the more complex loudness measure; however, when the drum sounds are more similar in character, loudness is a better measure. As loudness always shows good correlation to the perceived loudness, it is most reliable to use loudness in all cases.

The loudness measure has been used in several investigations of building acoustics. In 1965, Fasold used the loudness level when deriving an optimum theoretical curve for the normal impact sound level [13]. Psychoacoustic measures were used by Zwicker and

Fastl [14] to obtain quantitative figures regarding transmission loss for a facade. Good correlation to subjective preferences with regard to impact sound insulation was found by Nilsson and Hammer [15], using the standard tapping machine and the ISO 532B loudness measure. The use of loudness in standards is, however, limited, no such use being known to the authors. On the other hand, A-weighted and C-weighted sound pressure levels are used in many standards. A-weighted sound pressure level was standardised only a couple of years earlier than loudness was, but since it was easier to calculate by hand than loudness was, and could already be measured using a portable level meter in 1936 (loudness could first be measured using a portable meter only in 1981) it has been more widely used. With the advent of modern computers, however, these differences are less important. A-, B and C-weighted sound-pressure levels are derived from the 40-, 70- and 90-phon contours, respectively. As a result, these measures are applicable to certain sound levels, and predictions of the perceived loudness of sounds of various frequency and level contents based on one of these measures are likely to fail. The loudness sensation, however, depends not only on the frequency content; other effects, such as masking and sound duration, have to be considered as well. The loudness measure takes account of the level and frequency dependency of the ear by using several phon contours, the effects of masking, and spectral distribution [12]. If one standard is sought for several types of floor coverings, it is therefore better to use a measure that can adequately handle various levels of drum sound.

Two methods for calculating loudness are described in ISO 532. Part A is based on octave-band and part B on one-third-octave–band measurements. Although these methods were developed for use with steady sounds, part B was shown to be useable for transient sounds as well [4]. Loudness, unlike A-weighted sound pressure level, is a linear measure, i.e., a doubling of the loudness value indicates a doubling of the perception of loudness. This linearity is advantageous, particularly in communicating differences and improvements to non-acousticians, and some examples are given in [16].

The use of loudness to predict either perceived loudness or perceived disturbance is therefore suggested.

2.2 Excitation source – tapping machine

As the ISO tapping machine is used in measuring the impact sound of floorings and it produces the drum sound when operating, it seems natural to use it in measuring the drum sound as well. However, several other excitation sources come to mind. A walking person definitely produces the drum sound for that particular person, but the method naturally lacks the reproducibility required of a standardised method. A machine that simulates an actual, particular foot can be produced, but such a machine would not be immediately available to acousticians and manufacturers of floor coverings. The annex of ISO/DIS 140-11 [17] describes a modified tapping machine. The modification involves inserting a soft layer between the hammers and the floor surface to make the impedance spectrum of the hammer resemble that of a person walking without shoes. Scholl suggested a similar solution in [18], but the question remains as to whether this would be applicable for evaluating the drum sound. The soft layer was used in [6] in measuring drum sound, and it was seen that the material used was too thick because all results were the same and what was measured consisted primarily of inherent mechanical noise. This was hardly surprising, since the difference between most floor coverings in terms of the sound excited by a person walking without shoes is small compared to the difference when a person is walking with hard-heeled shoes. Therefore the modified tapping machine is not applicable here; the correlation to the real situation is instead achieved by using a proper method for evaluating what is measured with the original tapping machine available in most building acoustic laboratories. Another sound source that might be useful is a steel ball; however, measuring the sound of a single impact is more complicated than measuring sound from a repetitive, stationary source.

In this study the correlation between the measurement method and the subjective perception of the drum sound is always controlled for using a walking person wearing hard-heeled shoes as the exciter of the drum sound. However, drum sound is also produced by dropped items such as keys. The evaluation method presented here should probably be changed to assess accurately the drum sound produced by such items. On the other hand, the measurement set-up could stay the same, as long as the source excites the frequencies of interest. No investigations of this kind of drum sound will, however, be presented here.

The ISO tapping machine is described in annex A of EN ISO140-8 [8] which specifies, for example, the hammers' curvature, dimensions, location and momentum, as well as the time between impacts. Still, when two different tapping machines are used in testing the same flooring in the same room, discrepancies in the results may still appear. Assuming

that the requirements specified in EN ISO140-8 [8] are met, the following may explain such discrepancies:

- Inherent mechanical noise affecting the measured drum sound
- Diverse directivity of the excited drum sound
- Varying contact between the floor covering and sub floor: if exactly the same spot was not used, or if too few spots were used in the averaging process, divergent results will be produced. This is, however, not due to the tapping machine itself and will be treated in section 2.3.

As the tapping machine was developed to produce impact sound, its inherent mechanical noise was not regulated in the standard; such noise can be reduced when needed by installing a cover, without affecting the transmitted impact sound. When it comes to measuring drum sound, however, the inherent mechanical noise is more significant and not as easily dealt with. Figure 1 shows the results of measurements of drum sound made on two floors using two different types of tapping machines. The sound levels were measured at a distance of 1 m from where the hammers hit the floor surface in a damped room (reverberation time 0.2 s, 300−6300 Hz, room volume 162 m^3). The inherent mechanical noise clearly affects the results for the flooring in the right-hand diagram (Figure 1 b), while it has a lesser effect on the flooring in the left-hand diagram (Figure 1 a), although differences do still exist. To enable drum sound to be measured on all types of flooring, including textile flooring, the inherent noise needs to be decreased significantly. However, all of the currently available tapping machines fulfilling ISO 140-8 can not easily be improved enough, for example, by installing a cover, so as sufficiently to reduce the mechanical sound. The inherent noise needs to be measured in some way so that proper correction can be made, or at least so that it can be estimated whether the measurements are affected by the inherent noise.

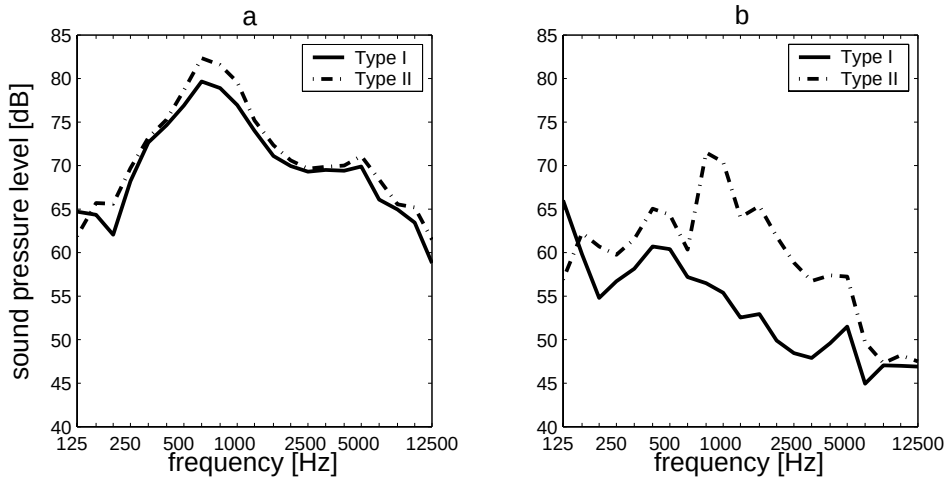

Figure 1 Measurements of drum sound using two types of tapping machines (types I and II) on two floor coverings: a) 7-mm veneer floor on polyethylene foam on concrete sub floor; and b) 3-mm linoleum flooring with a soft underlay glued to the concrete sub floor.

When measuring the inherent mechanical noise of the tapping machine it is important to simulate the actual situation as much as possible, although the true drum sound should be removed. Preferably the tapping machine should stand as stably as in the real test situation, soft material shall only be installed under the hammers, the correct falling height must be maintained and the distance between the tapping machine and the floor surface must be unchanged so that the radiation pattern is similar to that occurring during drum sound measurement (see Figure 2). The same figure shows the results of three different set-ups for measuring inherent mechanical noise. The results obtained with solid blocks of high-density fibreboard (HDF) and concrete blocks, respectively, and when the entire tapping machine is standing on a textile carpet are shown. Measurements are made at a distance of one meter from the sound source in a damped room, as described above. It can be seen in the figure that the amount of inherent noise differs between the three set-ups. The inherent mechanical noise of the tapping machine depends on the stability of the surface supporting the machine. The textile carpet introduces absorption and stabilises the machine, thus decreasing the measured inherent mechanical noise. If correction is to be made for inherent noise, it seems like none of these measurements is applicable for all types of drum sound measurements. As the inherent mechanical noise depends on the

whole system, a general specification on the measurement set-up could be to use solid blocks (as in Figure 2) of the same material as the actual test floor.

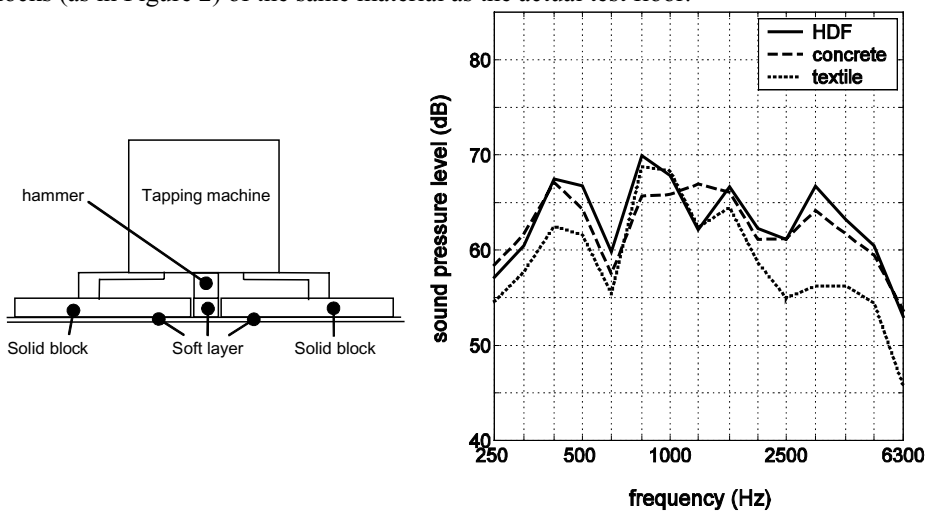

Figure 2 Measurement of inherent mechanical noise using three set-ups with the same tapping machine. The results obtained using the set-up to the left with solid HDF or concrete blocks is shown by the solid and dashed line respectively. The dotted line is obtained when placing the entire tapping machine on a textile carpet.

In the absence of new tapping machines on the market, developed to enable drum sound measurements as well, there is a need to find a way to reduce inherent mechanical noise. One such way could be to use a sound-reduction cover that muffles the inherent sound without affecting the radiated drum sound. The French standard [7] suggests a cover of either glass or stone wool. Unfortunately, this solution is not good enough for all types of tapping machines. The newer available tapping machines have too much inherent noise for this solution to be effective, even for measurements made on thin floating floors such as veneer and laminates. If the cover is made of a rigid material lined with a porous absorber, the sound-reduction effect increases but more of the emitted drum sound is also muffled. As the intention of the cover is to reduce the mechanical noise without affecting the emitted drum sound, the dimensions of the cover should be in the same range as those of the tapping machine. The cover should be designed to be most effective at frequencies at which inherent noise is greatest. A rigid cover could well decrease the measured sound

at some frequencies while increasing the measured sound at others, due to internal resonances of the cover at frequencies of about 200 Hz and lower, depending on the geometry of the cover. As the sound pressure levels at these frequencies in most cases, for example, involving thin floating floors on a homogenous concrete sub floor, have less effect on the perception of the drum sound, they could be neglected. However, for some floor coverings, such as 22-mm wooden flooring on timber joists, there is in general a great amount of energy at these frequencies, so neglecting them might lead to misleading results. Therefore, the design of the sound-reduction cover should be carefully examined so that proper measurements of the drum sounds of interest are enabled.

Larris [5] suggests another way to reduce inherent mechanical noise: reducing the impact frequency of the steel hammers. If the design of the tapping machine is not stable, the movements of the five hammers interact so that the machine starts to vibrate even more. For impact sound five hammers are needed to obtain sufficient power; for drum sound, however, such high power is not needed. The decrease of inherent noise obtained by using one rather than five hammers must be compared with the decrease of drum sound resulting from the same change. If the decrease of inherent noise is greater than the decrease of drum sound, the signal-to-noise ratio increases, resulting in overall improvement. If no difference is measurable it can be argued that it is better to use five hammers, since noise other than the inherent noise will be better masked. Figure 3 shows the difference between using five hammers and one hammer (the middle hammer) for two types of tapping machines. The difference in the sound pressure level of the inherent noise obtained with five as opposed to one hammer is subtracted from the difference obtained in the sound pressure level of the drum sound obtained with five as opposed to one hammer as measured on a laminate floor covering. Hence, a positive difference indicates that the difference between the inherent noise and drum sound has increased, while a zero result means that no gain or loss has been achieved by using one hammer instead of five. Measurements are made at a distance of one meter from the sound source in a damped room, as described above. All results represents median values (four measurements were made of the inherent noise, eight of the drum sound). As can be seen in the figure, for the type I machine there are positive and negative differences, while the for type II machine – also the machine producing more inherent noise – the difference is positive for most one-third-octave bands. As the difference between the inherent noise and drum sound for this type of tapping machine can be as small as a few decibels, the difference of about 4 dB that is seen in the 500–4000 Hz range is helpful. However, using only one hammer increases the standard deviation of the measurements, since the drum sounds are created at fewer spots. In these measurements the median one-third-octave–band standard deviation was 2.0 dB for five hammers but 2.8 dB for one hammer.

Yet, the use of only one hammer could be one way to decrease the inherent noise of a tapping machine, although the amount of decrease depends on the design of the machine.

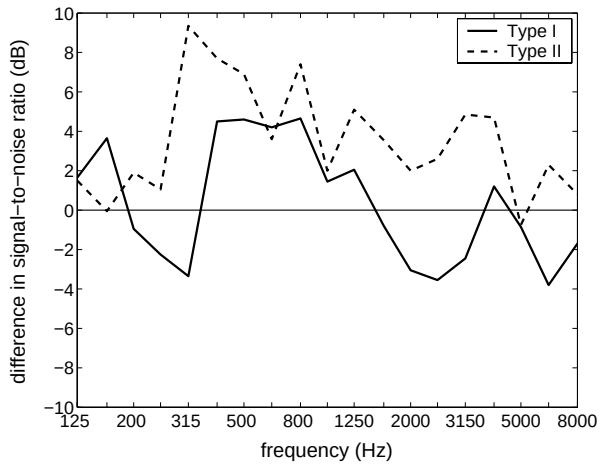

Figure 3 Difference in signal-to-noise ratio using one hammer and five hammers. The signal is the measured drum sound with an 8-mm laminate floor covering on a concrete sub floor; the noise is the measured inherent noise. The difference is measured for two types of tapping machines, I and II.

The weight of different types of tapping machines varies. Different weights alter the contact between the floor covering and the sub floor, and this could have an effect on the emitted drum sound. Therefore the total weight of the tapping machine should be specified.

The directivity of the two types of tapping machine for two, one-third-octave bands is shown in Figure 4, zero degrees denoting the case where the axis of the hammers is in line with that of the microphone. At 630 Hz, there is a clear dependence on the angle, whereas at 1000 Hz the pattern of radiation from the tapping machine is more similar to a half sphere. The frequency dependency of the directivity of the two types of tapping machines is similar, although one shows greater difference between 90 and −90 degrees due to the asymmetric construction of the tapping machine. Adequate power measurements can be made as long as the same total power is emitted from the machines. As measurements made at one point in the direct field will not produce the same results for all points using different tapping machines, a control is needed indicating that the selected measurement points produce similar results. The directional properties of the

tapping machines mean that it is important to remove any reflecting objects close to the measurement set-up.

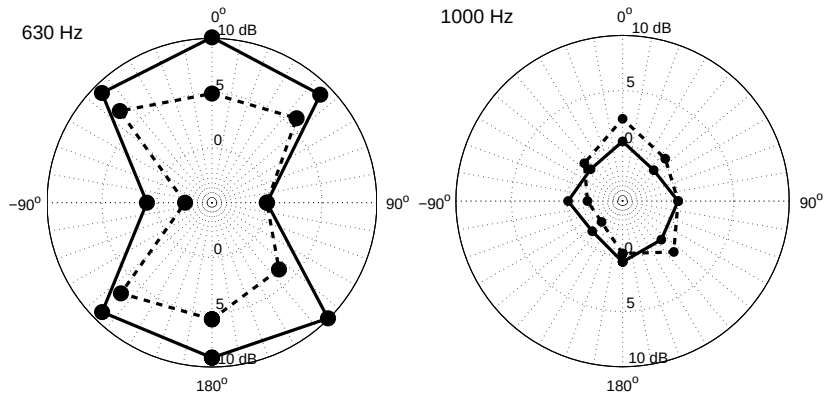

Figure 4 Directivity of two types of tapping machines, type I (dashed line) and type II (solid line). Zero degrees denotes the case where the axis of the hammers is in line with that of the microphone.

Even though there are difficulties in reducing the inherent noise, finding a proper method to measure inherent noise is crucial. It is also important to take account of the fact that different types of tapping machines can produce somewhat different directivity patterns. The tapping machine can be used as an excitation source for the drum sound, to rank hard floor coverings, such as laminate, veneer and parquet flooring; it should, however, be used with caution for drum sounds of low levels.

2.3 Influence of uneven contact between floor covering and sub floor

When making measurements on a floor covering over a thin hard underlay material the resulting sound pressure level can vary by 10 dB over a one-third-octave band just by moving the tapping machine to another spot – even while keeping the tapping machine oriented in the same direction relative to and the same distance from the microphone (see Figure 5). This difference is seen even though the floor fulfils the demand in [8]: in height ±1 mm over a horizontal distance of 200 mm. If the exact same spot was not used or if an insufficient number of spots was used in averaging, different results will be obtained. The flooring does not, on the other hand, seem to be as sensitive to air gaps when a real foot is creating the drum sound. The foot apparently presses the floor covering to the sub floor and removes the air gap before the major part of the sound is radiated. It was observed that for a floor covering with a thin (< 1 mm) underlay on a

concrete sub floor, positioning the tapping machine so that the influence of air gaps was small corresponded better to subjective perceptions of the sound of a person walking on the floor covering.

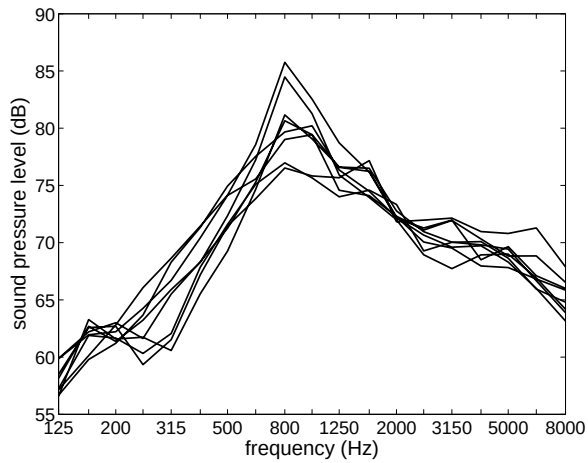

Figure 5 The lines represent drum sound measurements made on 7-mm laminate floor covering with a thin, 0.3-mm underlay using a tapping machine in eight different positions. Uneven contact between the floor covering and the sub floor may cause the radiated drum sound to vary 10 dB in a one-third-octave band.

As a result, the method chosen should take this effect into account, to minimise the influence of any air gaps.

2.4 Influence of the size of the test floor

Changes in the dimensions of the test sample could influence the radiated drum sound. A resonant floor covering with little damping is more sensitive to such changes than is a locally reacting floor covering with high damping. From a practical point of view, a small test floor is preferable since it is easier to find a suitable small test room, less work is required to install the floor and less material is needed. However, it remains to be verified that the radiated drum sound is not influenced by the reduced size of the floor. When testing the sound reduction impact of floor coverings on a heavyweight floor according to ISO 140-8 [9], the size of the floor covering shall be at least 10 m^2 with the smaller dimension being at least 2.3 m. A floor covering of 8-mm laminate (high-pressure laminate + HDF) with an attached 2.5 mm polyurethane underlay material was used in a test where the floor size was decreased from 3×4.1 m^2 to 2×2.4 m^2. Eight different

positions of the tapping machine on the floor were measured using 4 different microphone positions (see section 3.2) on both test samples. The measured sound pressure levels are shown in Figure 6. A two-sample *t*-test [19] comparing the mean value in each one-third-octave band revealed no systematic changes resulting from the changed size of the test floor at the 0.05 level of significance.

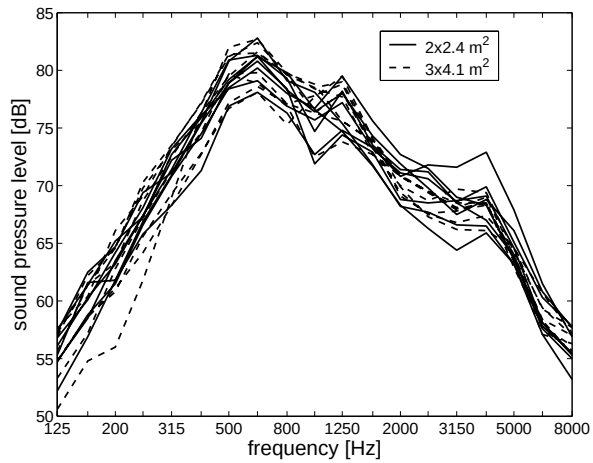

Figure 6 The radiated drum sound from two samples of different sizes was measured. The floor covering was an 8-mm laminate with an attached 2.5 mm of polyurethane underlay material.

Thus, for this type of flooring the smaller floor size (2×2.4 m^2) is sufficient. Naturally, other types of flooring might display different behaviours that need to be investigated. However, the flooring dimensions should preferably be chosen so that the ratio of the squares of the edge lengths is irrational, so as to excite the most eigenmodes of the flooring [20]. Here the focus has been on floor coverings; no investigation has, for example, been made of the influence of test floor size on lightweight floor structures.

2.5 Test room

The airborne and impact sound reduction properties of floor coverings are tested in laboratory according to ISO 140-8 with the use of two adjacent reverberation rooms. Measuring the drum sound in such an environment permits the rapid and convenient measurement of these three properties. However, floor manufacturers are not always interested in all three properties. Good drum sound properties generally mean poorer impact sound properties and vice versa, and manufacturers are sometimes interested in a product with the best available drum sound properties, regardless as to its airborne or

impact sound properties. A reverberation room with a diffuse sound field is available in most sound laboratories, but such conditions are not easily achieved by manufacturers wishing to make the measurements themselves. A procedure that enables measurements to be made in the field is much sought after. Field measurements made using an evaluation method relying on a diffuse sound field have the disadvantage that a diffuse sound field generally never exists. However, a sound power level measurement method similar to EN ISO 3744 [21], in which a damped room is required, could be a solution.

As previously stated, an important requirement of any measurement procedure is that it should produce a measure with a subjective counterpart. As the perception of a sound depends on, among other matters, the level and frequency content of the sound, correlation to the subjective perception is likely to improve if the level of the measured drum sound is equal to that of the original sound. A damped room is more similar to an office space or home environment than is a reverberation room. Measurements made in a more damped room indicate lower sound pressure levels than do measurements made in a reverberation room, and are more in accordance with actual experience (even though the level of the drum sound produced by the tapping machines still exceeds the level produced by the average walker).

Figure 7 shows the correlation between a listening test of the perceived loudness of a drum sound generated by walking persons and four measures of the sound pressure levels measured at various laboratories with different reverberation times. The results were obtained from the round-robin study described below in section 3.5. The sound pressure levels were measured eight times at a distance of 1 meter from the tapping machine at a height of 0.71 m, as in EPLF NORM 021029-3 [1]. The median values of the calculated measures for each of the five floor types were used to calculate the correlation coefficient for the listening test. It is seen in the figure that the loudness measure gives the best correlation for reverberation times less than 0.4 s, but that for the two laboratories with reverberation times of 1.0 and 1.6 s, respectively, the correlation coefficient is low. A reason for the decreased correlation could be the directivity of the tapping machine. As was seen in Figure 4, for some frequency bands the tapping machine radiates more sound in directions other than the measurement direction. When the reverberation time is increased, the influence of the reverberant field increases, which could cause the changes in the spectra energy distribution seen in Figure 8. The measurements are parallel, but the curve of the longer reverberation times shows a second peak at 5000 Hz. This difference might seem small, but as the differences between the examined floor coverings were rather small it is sufficient to decrease the correlation to the listening results. It seems as if the measurement set-up used in EPLF NORM 021029-3 [1] cannot be used in rooms with longer reverberation times. However, with a different set-up, measurements can be

made in rooms with longer reverberation times. In Johansson and Nilsson [11] laboratory measurements were made in both a damped ($T_{250\text{-}2000}$ = 0.2 s, volume 162 m^3) and a reverberation room. The measurements in the reverberation room were made using a rotating microphone and several types of tapping machine. Independent of the measure used (loudness, A- and C-weighted sound pressure levels), the correlation coefficients were always somewhat higher for the damped room, though still of a similar magnitude (all correlation coefficients were between 0.80 and 0.89). However, measurements made in a reverberant field display another effect that needs to be considered. The coupling between the floor and room can in adverse circumstances lead to misleading results, especially in the low-frequency region, as what is measured is not only the properties of the floor but also the properties of the room.

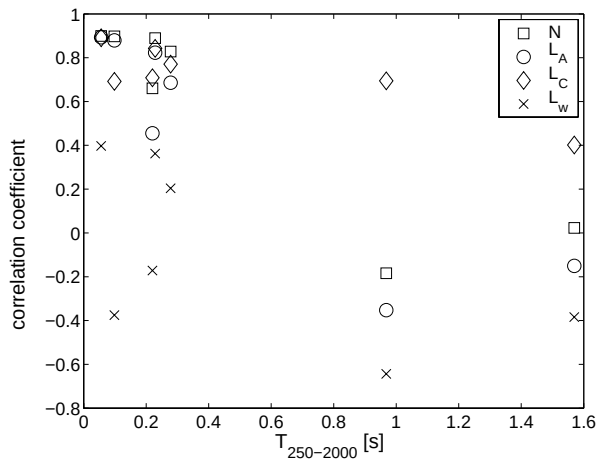

Figure 7 The correlation coefficients of the results from the round-robin test conducted in seven laboratories with various reverberation times, $T_{250–2000}$, to the listening test described below are calculated using four measures: loudness, N, A- and C-weighted sound pressure levels, L_A and L_C, and weighted impact sound pressure level, L_w.

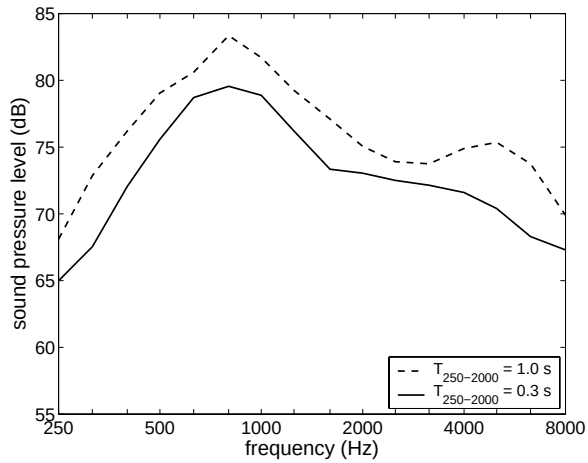

Figure 8 The median of eight measurements of drum sound for a laminate floor covering with cross-linked polyethylene underlay foam at two locations having different reverberation times.

As in the case of impact sound reduction measurements, the influence of a floor covering on the drum sound is dependent on the sub floor. Measurement results for a bare, reinforced concrete slab will be dependent on its thickness, but when a floor covering such as thin wooden parquet is added, this dependence is reduced as the difference in impedance is large. However, if all types of floor covering are to be tested using the same method, and if limits on the drum sound of floor coverings are to be set, a demand must be established as to the thickness of the sub floor so that its influence is reduced or its effect accounted for (see [7]).

Measurements of drum sound can be made in a damped or a reverberation room: the choice is not critical, though the measurement set-up needs to be adapted to each situation so as to ensure good correlation to listening test results. However, the final results obtained in a damped room have been shown to agree somewhat better with subjective perceptions of drum sound.

3 BRANCH NORM – EPLF NORM 021029-3

3.1 Defining the scope of the norm and choosing initial method

In 2001 the branch organisation, European Producers of Laminate Flooring (EPLF), started searching for a method for measuring drum sound on laminate floor coverings. EPLF required that the method should be repeatable, reproducible, practical and must produce results in line with subjective perceptions of the sound. There was also a desire to find a method that could be performed without a reverberation room.

Initial listening tests and objective measurements were performed for four laminate floor coverings. The tests used the following exciters of the drum sound: a dropping ball (Wilhelm-Klauditz-Institute, WKI, in Braunschweig, Germany), an actual foot (Institut für Holztechnologie, IHD, in Dresden, Germany), an ISO standard tapping machine in a reverberation room (Centre Technique du Bois et de l'Ameublement, CTBA, in Bordeaux, France) and an ISO standard tapping machine in a room with a short reverberation time (Engineering Acoustics, Lund University, in Lund, Sweden). The results of all methods were observed to correspond well in terms of perceived loudness with the results of listening tests; this was true when the results were either in their initial form or when loudness according to ISO 532B was used to evaluate the measurement results. The listening tests included questions about the pleasantness and pitch of the sound. Good correlation to 'pleasantness' was also achieved, but this result was excluded since what is perceived as pleasant in some cases includes a judgement as to whether the sound is deemed appropriate for the type of the floor (i.e., wood, stone, etc.). Hence, a sound that is judged to be pleasant for a wood floor, for example, might not be perceived as pleasant when emanating from a stone floor; the perceived loudness, however, is not influenced. When the design or type of floor is unknown to the listener, such as when the listener is listening to recorded drum sound, the perceived loudness and perceived disturbance are highly correlated [11]. No measure with good correlation to perceived pitch was found, and it was finally decided to focus on finding a suitable norm method that correlated well to the perceived loudness of the drum sound produced by walking persons.

As the method presented by the Division of Engineering Acoustics, Lund University, allowed measurements to be made in rooms that were more available to the branch and used the tapping machine already used in impact sound measurements of floors and floor coverings, it was chosen as the initial method. However, the method has been revised since then, the final revision being made in November 2004 after a final round robin

including listening tests was finished [1]. A brief description of the criteria and limits of the norm follows.

3.2 Measurement set-up

The ISO standard tapping machine is chosen as the source of the drum sound. A sound-reducing rigid cover lined with sound-absorbing material is used to reduce the inherent mechanical noise of the tapping machine itself. The sound-absorbing material was selected to function optimally in the 800–1000 Hz range where the inherent noise is greatest. Frequencies below 250 Hz are neglected, since their influence on laminate floorings is small and since considering them would place demands on cover design that would be difficult for different types of tapping machines to accommodate. Since differences were observed between types of tapping machines, one type of tapping machine is selected or tapping machines proven to produce identical results can be used. The type chosen is not the one with the lowest inherent mechanical noise, but it is a type that is readily available. Inherent noise is measured to get an indication of whether or not the results are affected by inherent noise; no corrections are made, however, since more information is needed concerning how to make such measurements properly.

Since the flooring industry would prefer a method that could be performed without a reverberation room and preferably in the field, a room with a short reverberation time is chosen as the test room. Hence, a room with a free field over a reflecting plane, as in a semi-anechoic room, is desirable though not required. Such a room was also chosen as it is more similar to a real-life environment, such as an office space, than a reverberation room is. To make complete power measurements of the drum sound produced by the tapping machine, microphone positions must be chosen so that a representative value of the sound power is achieved. EN ISO 3744 [21] requires at least eight microphone positions. However, the process of measuring drum sound is complicated by varying contact between the floor surface and the sub-floor, and several measurement positions on the floor are needed in order to achieve representative results; truly representative sound power measurement thus becomes time consuming. Instead, a fixed position of the tapping machine relative to a single microphone is chosen. Eight different positions of the tapping machine on the floor covering relative to four different microphone positions is used (the tapping machine is moved twice while retaining the microphone position). To decrease the influence of the reverberant field on the measurements, a distance of one meter between the microphone and where the centre hammer hits the floor is chosen. When measuring in the field it is difficult to decide whether measurement is being made in the direct or in the reverberation field. Departures of the test environment from free-field qualification are partly accounted for, as in EN ISO 3744 [21]. The measured sound pressure levels are corrected with the environmental correction factor, K, defined as

$$K = 10\log(1+156.1\cdot\frac{T}{V})\ \text{(dB)}$$

where T is the reverberation time (s) for the actual one-third-octave-band and V is the volume of the room (m^3). The free-field qualification is satisfied by a given test room if the ratio of the volume of the room to the reverberation time is sufficiently small; in EN ISO 3744 [21] K must be under 2 dB. It was seen in the round-robin test that the demand on the quoted T/V needed to be raised to ensure measurements in the direct field; an upper limit on T of 0.45 s in each one-third-octave-band was chosen together with the demand that the correction factor should be less than 2 dB, as in EN ISO 3744 [21]. As measurements of reverberation time less than 0.15 s are insecure it was chosen as the lower limit in performing correction, even though measurements are allowed.

The size of the tested floor covering shall be 2×2.4 m^2; the sub floor shall consist of reinforced concrete at least 120 mm thick.

3.3 Evaluation

Loudness according to ISO 532B is chosen, as it has been shown to be superior to other suggested measures, such as A-weighted sound pressure level, as reported in section 2.1. It is assumed that the four lowest loudness values represent positions where the influence of any gap between floor surface and sub floor is small, and good contact is achieved (see section 2.3). The mean of these four values is then taken to represent the floor covering and is denoted N_m. This N_m value is then compared with the N_m value of the reference, so as to obtain the reduction or increase of the drum sound's loudness in percent.

3.4 Classification

As reported above, a measure that correlates well to the perception of the drum sound is presented in Johansson et al. [4]. The goal of the study was also to gain understanding of the subjective implications of differences in these objective measures; that is, what values of a given objective measure align with different qualitative judgements. In a paired comparison test performed in the laboratory, listeners were asked which walking sound was the most disturbing. The responses were analysed using a modified Bradley and Terry model allowing for ties [22]. Various measures, such as loudness according to ISO 532B, were tested against the subjective response using linear regression. The difference in 10-percentile loudness, N_{10}, between two stimuli was shown to predict the subjective response better than, for example, A-weighted sound pressure level. A difference of about 8% in N_{10} or 8% in loudness resulted in 50% of the subjects declaring a qualitative difference.

Within the norm, a reference floor covering was used to establish a basis of comparison. The worst class consists of floor coverings producing increased loudness compared with the reference or loudness reduction less than 5%; the next better class produces loudness reduction of 5 to 15%, the next class 15 to 25%, etc. These categories were judged appropriate based on the findings of the tests in the round-robin study and the test reported above.

3.5 Round-robin study and listening test

To finalise the work on the norm it was decided to perform an extended interlaboratory round-robin study. The main objectives of the study are to determine and document the within-laboratory repeatability and the between-laboratory reproducibility of the test method and to determine and document the correlation between the objective results of the test method and the results of subjective listening tests.

Five laminate floor coverings are included. Two of the laminate floor coverings consisted of 7-mm HDF: one was installed loosely over a layer of polyethylene foam and the other over a soft board. The three other laminate floor coverings incorporated attached underlays of polyurethane foam, cross-linked polyethylene foam and thin (0.3 mm) underlay, respectively. The floorings are referred to as flooring A, B, C, D and E, the order of labelling being randomised so as to anonymise the results.

Round robin

The laboratories participating in the round robin are:

- Alveo AG, Luzern, Switzerland
- Danish Electronics, Light and Acoustics, DELTA, Copenhagen, Denmark.
- Engineering Acoustics, LTH, Lund University, Lund, Sweden
- Institut für Holztechnologie in Dresden, ihd, Dresden, Germany
- SP, Swedish National Testing and Research Institute, Borås, Sweden
- Wilhelm-Klauditz-Institute, WKI, Braunschweig, Germany

The measurements in the round-robin study were made according to EPLF NORM 021029-2 [23], a former version of EPLF NORM 021029-3. The major differences between the two norms are that in the former version the frequency interval of interest is larger, the environmental correction somewhat different, no reference is used and no loading of the test samples is conducted. Each laboratory measured the drum sound twice. While evaluating the results, however, revisions were made. The values reported here are based on these measurements, but the evaluation of the measured sound pressure levels has been adapted according to EPLF NORM 021029-3, as follows:

- Frequency interval of interest is 250–6300 Hz
- Environmental correction is made if the reverberation time is 0.15–0.45 s. When the reverberation time is less than 0.15 s no correction shall be made, although measurements are allowed. For reverberation times greater than 0.45 s no measurements are allowed. The measurements made at DELTA, where the reverberation time was long (up to 1.1 s), would not be sufficiently corrected by the environmental correction factor, so their results are not presented here.
- The 7-mm HDF with polyethylene underlay foam is used as a reference (in result section 'B'), and the results for the four other floorings are presented here in terms of percentage reduction compared with the reference loudness value, N_m. The reference mentioned above in section 3.3 was chosen after the round robin.
- One type of tapping machine and other tapping machines producing identical results are allowed. As two of the labs used another type of tapping machine producing different results, they were excluded from the evaluation below.

The laboratories are denoted Labs 1 to 4, the order not corresponding to that of the above list.

Listening test

The listening test was performed by the Division of Engineering Acoustics, Lund University, Sweden. The sounds were recorded in a room of 162 m^3 having a reverberation time of 0.3 s at 200–315 Hz and 0.2 s at 400–10000 Hz. The level of the background noise (slight fan noise) was 21 dBA. The samples were laid on the bare concrete floor in the middle of the room (see Figure 9**)**. Four walkers walked, one at a time, back and forth taking five steps in each direction over the sample at a speed of approximately 2 steps/s for 30 s. Two male and two female walkers wearing hard-heeled shoes were used. Two-channel recordings of the walking sounds were made using a dummy head and B&K 4181 microphones, G.R.A.S. 26AK preamplifiers and the 01-dB Stell Symphonie digital real-time analyser (sampling frequency 51.2 kHz). The dummy's ears were located 1.15 m above the floor surface, as would be the case with a sitting person.

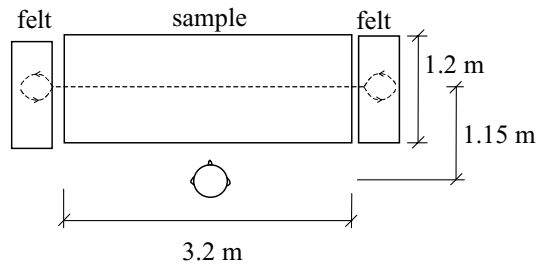

Figure 9 Measurement set-up for making recordings for the listening test.

From the 30 s recordings four samples were selected for inclusion in the test. The four first five-step sections available were selected unless irregularities in the walking, such as obvious stumbles, were perceived in any of them. The duration of each selection was 2.8 s.

During the listening tests the listeners listened to the recordings using Sennheiser HD 600 headphones. The playback level using headphones was checked by direct comparison of the live and recorded drum sound, this calibration being performed twice by five listeners: the agreement of the five was good, and the average value of the chosen playback level was used. The playback level was tested in a second way. White noise was recorded in the room and then replayed through the headphones on the dummy head. The playback level was adjusted and analysed until the same level as that of the initial white noise was achieved. The two methods resulted in the same playback level.

The listening test was performed in a conference room of a laminate company in Perstorp, Sweden, and at the Division of Engineering Acoustics, Lund University. The background noise from the computer was minimised. The paired comparison test was performed using MATLAB. The listeners were introduced to the test and familiarised with the sounds before the test started, and a preliminary example test was conducted with no collection of results. The listeners were thereafter requested to imagine themselves in an office space, and in a paired comparison test, to say which of the sounds was louder (Swedish, *ljudstarkast*). It was possible to declare a tie. The subjects could switch between the sounds as many times as desired, with no time limit. As a button was pushed, a sound sample of five steps was played. The sound sample was chosen randomly from the available five-step intervals from the original 30 s recordings.

The order of the sound samples was randomised for each subject. Each listener compared each pair of sound samples once; the length of the session was approximately 15 min.

During the test, the listeners were left alone, although they could get the attention of the test leader if necessary. The test results were saved automatically in data files for further analysis.

Fourteen female and eighteen male listeners participated, their ages ranging from 25 to 61 years. None of the subjects reported any hearing disabilities. Scale values for the floorings were produced using the statistical paired comparison model of Rao and Kupper [22]. This model allows ties, assuming that when the difference between two treatments is less than a certain value, or threshold, subjects will declare a tie. The probability of choosing treatment T_i when compared to T_j is set to

$$P(T_i \rightarrow T_j) = \int_{-(\ln \pi_i - \ln \pi_j)+\ln \theta}^{\infty} \operatorname{sech}^2(y/2)\,dy = \frac{\pi_i}{\pi_i + \theta \pi_j}$$

where π_i (i = 1, 2…t) represents probability values for the t treatments, where $\pi_i \geq 0$ and $\sum_{i=1}^{t} \pi_i = 1$. $\ln(\theta)$ is the sensory threshold for the subject. $\ln \pi_i$ is regarded as the 'true' merit of treatment T_i, and the natural logarithms of the probability values form a linear scale rating of the treatments. In this research, a higher value indicates a higher perceived loudness.

Results

Figure 10 presents the results from each laboratory. The reduction (%) is calculated using the loudness value N_m of floor covering B as the reference value. Table 1 presents the median values from the four laboratories; the medians of the following measures are also shown: loudness, A- and C-weighted sound pressure level, and weighted impact sound pressure level. The weighted impact sound pressure level was calculated using the reference curve in EN ISO 717-2 [9]. All values are calculated using environmentally corrected sound pressure levels. The same table presents the scale values for the five different floorings. Scale values for each walker are calculated from the listening test (see Figure 11). The variance of the scale values for each walker was small, but different walkers produced slightly different rankings (between floorings A, C and D). The table shows the median values for the four walkers.

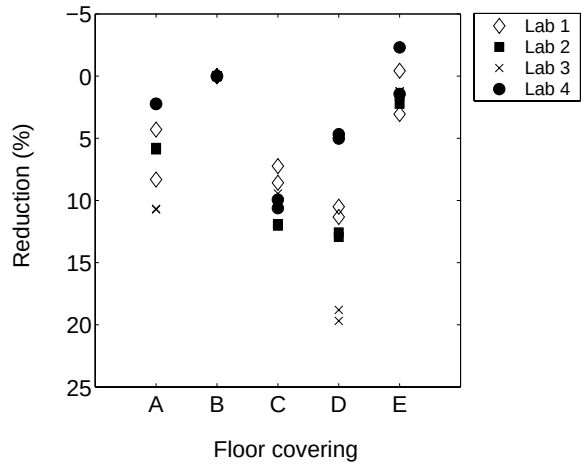

Figure 10 Results of the round-robin study. Each laboratory performed the measurements twice. The loudness of floor covering B is used as the reference when calculating the reduction (%) of the drum sound's loudness.

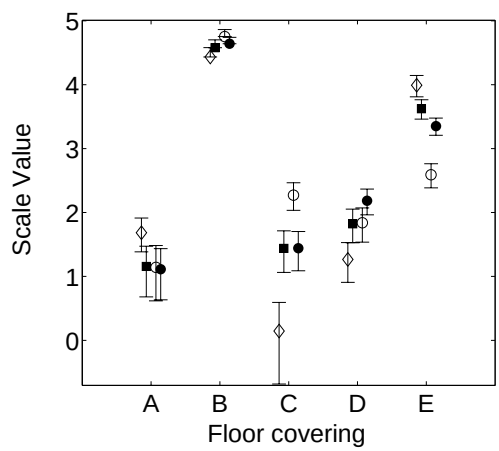

Figure 11 Scale values from the listening test together with their 95% confidence intervals. The open circle, filled circle, open diamond and filled square represent the four walking persons.

Table 1 The median results from the four laboratories for the following measures: reduction compared to floor covering B (%); loudness, N (sone); A- and C-weighted sound pressure level, L_A, L_C (dB); and weighted impact sound pressure level, L_w (dB). The scale values given are the median values for the four walkers. The coefficients of correlation between the measures and the scale values are given.

Floor covering	Reduction (%)	N (sone)	L_A (dBA)	L_C (dBC)	L_w (dB)	Scale values
A	5.8	60.9	86.1	86.4	80	1.2
B	0	66.4	86.8	88.5	80	4.6
C	10.2	60.5	85.9	87.1	79	1.5
D	12.8	56.9	85.1	85.2	79	1.9
E	1.4	64.6	86.4	90.0	79	3.5
Correlation to scale values	–0.80	0.83	0.72	0.73	0.16	

The precision of the test was determined with the help of ASTM E691 [24] and the results are shown in Table 2. Floor covering B is excluded since it is used as a reference. S_r is the within-laboratory standard deviation of the mean and repeatability, $r = 2.8S_r$. S_R is the between-laboratory standard deviation of the mean and reproducibility, $R = 2.8S_R$. In comparing two mean values for the same material obtained by the same operator using the same equipment on the same day, the means should be judged not equivalent if they differ by more than the r value for that material and condition. In comparing two mean values for the same material obtained by different operators using different equipment on different days, the means should be judged not equivalent if they differ by more than the R value for that material and condition. The judgements of repeatability and reproducibility will have approximately a 95% probability of being correct.

Table 2 Precision statement of the round robin

Floor covering	Average (%)	S_r (%)	S_R (%)	r (%)	R (%)
A	6.3	1.4	3.6	4.0	10.1
C	10.0	0.6	1.7	1.8	4.8
D	12.0	0.5	5.9	1.3	16.6
E	1.0	1.8	1.8	5.1	5.1

Discussion

Figure 10 and Table 2 show that, compared with the differences between the floor coverings, repeatability is good but reproducibility is not as good. The results for two floor coverings from two different laboratories need to differ by 5 to 17 percentage points (floor covering dependent) to display significant differences. Floor covering D showed the lowest reproducibility; this, however, was expected since its underlay material was the thinnest. To be able to say whether or not this reproducibility is acceptable, a comparison is made using the reproducibility values regarding impact sound in ISO 140-2 [25]. The measurement environment specified in the EPLF norm is not as strict as that used in the laboratory measurements, but is more strict than that specified for field measurements in ISO 140; in this respect the reproducibility values obtained should lie somewhere in between. As the construction of an entire floor structure includes more uncertainties than the set-up of a laminate floor covering does, the reproducibility values for drum sound here should be lower than for impact sound in ISO 140-2. However, as a rough estimate, the reproducibility values obtained from laboratory and field measurements, respectively, were added to the median values obtained from the four laboratories for all five floor coverings (as reproducibility values are only given up to 3150 Hz, the value at 3150 Hz is used for frequencies above 3150 Hz). The loudness values are calculated based on these results and the original median values, and their differences are calculated. The difference is approximately 12 sones or 25–30 sones, using the laboratory and field values, respectively. As the difference between the results for a floor obtained by the four laboratories is at most 11 sones, the reproducibility values are within the expected range, even though improvements are always desirable.

The correlation to the listening test is good but not perfect. However, of the measures presented in Table 1, loudness, either used alone or together with a reference as in the reduction index, shows the highest correlation. A- and C-weighted sound pressure levels both show lower correlations. Weighted impact sound pressure level, L_w, shows no correlation at all and seems to be too approximate a measure. As can be seen in Figure 11, use of different walking persons resulted in different rankings of floorings A, C and D. This also indicates that the differences tested here are sometimes rather small; thus, the fact that the objective results obtained are not fully in agreement with the listening test results should not be considered as overly significant. However, the results indicate the need to use several walking persons, more than were used here, if similar floor coverings are to be ranked.

The results obtained from the other two labs using the other type of tapping machine were as well correlated to the listening test as were the results used above; these results were, however, excluded since it was observed that they always differed slightly from the

others. However, before this difference could be examined further, the committee of the technical group within EPLF decided for the time being to allow only one type of tapping machine (other tapping machines are acceptable providing their results are identical to those of the selected type of machine).

3.6 Applicability to other types of floor coverings

The EPLF NORM 021029-2 has been tested in a Nordtest project using other floor coverings than laminate floor coverings to test the method's applicability [11]. It was shown that the method can be used as a tool in selecting and ranking hard floor coverings, such as laminate, veneer and parquet flooring, but should be used with caution for drum sounds of low levels.

4 CONCLUSIONS

The measure loudness best correlated to subjective perceptions of the loudness and disturbance of the drum sound. The loudness is important when the degree of disturbance is judged, but other matters, such as the character and duration of the sound, are also important, although measures describing them have not yet been found.

The standard tapping machine without an additional cover cannot be used on all types of flooring without restrictions, due to the inherent mechanical noise of the machine itself. Even with additional covers, not all tapping machines can be used. The mechanism of the tapping machine needs to be improved to enable the tapping machine to be used in measurements of all types of floor coverings. The use of only one steel hammer can decrease the inherent noise and improve the signal-to-noise ratio.

If the inherent mechanical noise is to be corrected for, a proper method for measuring it is needed. As the inherent noise is system dependent, the measurement method needs to be adapted to the type of floor covering tested.

Two different types of tapping machines may produce somewhat different drum sound results even though the influence of the inherent mechanical noise is negligible. In the case of measurements in the free field this could be due to different directivity patterns of the machines.

The tapping machine is sensitive to any air gaps between the floor covering and the sub floor. A measurement method which examines several measurement points is therefore needed.

Measurement of drum sound can be made in either a damped room or in a reverberation room; the choice is not critical, but the measurement set-up needs to be adjusted to the measurement location. However, the final results obtained with a damped room have shown somewhat better agreement with subjective perceptions of the drum sound.

The branch norm EPLF NORM 021029-3 can be used as a tool in selecting and ranking hard floor coverings, such as laminate, veneer and parquet floorings, but should be used with precaution for low-level drum sounds due to the inherent mechanical noise of the tapping machine itself. The round-robin study showed that the loudness measure according to ISO 532B is best correlated to perceived loudness as determined by the listening test. The reproducibility (between-laboratory) is high, considering the differences being measured between different materials, but still lies in the region expected in light of the reproducibility of impact sound measurements.

5 RECOMMENDATIONS

The use of the measure loudness is recommended as it is best correlated to the subjective perception of the drum sound regarding perceived loudness, independently of whether or not the source is the sound of a real step or a tapping machine.

Measurement of drum sound can be made in a damped room or in a reverberation room; the choice is not critical, but the measurement set-up needs to be adjusted to the measurement location. However, the final results obtained with a damped room have shown somewhat better agreement to subjective perceptions of the drum sound.

When measuring drum sound it is best to use a tapping machine with a low level of inherent mechanical noise. A specially designed cover to reduce the inherent noise is needed in most cases. The use of only one hammer can decrease the inherent noise and improve the signal-to-noise ratio. The departments developing tapping machines are thus requested to seek ways to reduce such inherent noise.

The method presented can be used as a tool in selecting and ranking hard floor coverings, such as laminate, veneer and parquet floorings; it should, however, be used with precaution for low-level drum sounds.

ACKNOWLEDGEMENT

The authors wish to acknowledge the help of the members of EPLF technical committee and of the institutions that collaborated in developing the EPLF NORM 021029-3 branch norm: Institut für Holztechnologie (ihd) in Dresden, Germany; Wilhelm-Klauditz-Institute (WKI) in Braunschweig, Germany; DELTA Danish Electronics, Light and Acoustics in Copenhagen, Denmark; and SP Swedish National Testing and Research Institute, Borås, Sweden. The authors are grateful to Robert Månsson who performed most of the measurements at the Division of Engineering Acoustics. The authors also wish to thank the Swedish Foundation for Knowledge and Competence Development and Pergo Europe AB for their financial support.

REFERENCES

[1] EPLF NORM 021029-3 *Laminate floor coverings – Determination of drum sound generated by means of a tapping machine*, EPLF, Association of European Producers of Laminate Flooring, Bielefeld, Germany, 2004

[2] A.-C. Johansson, E. Nilsson, *Determination of drum sound from laminate floorings*, Engineering Acoustics, LTH, Lund University, Lund, Sweden, TVBA 3117, 2002.

[3] ISO 532B:1975, *Acoustics – Method for calculating loudness level*.

[4] A.-C. Johansson, P. Hammer, E. Nilsson, Prediction of Subjective Response from Objective Measurements Applied to Walking Sound, *acta acustica ACUSTICA*, 2004, 90(1), 161-170.

[5] F. Larris, *Drum noise from floors*, Teknologisk Institut, Lydteknisk Konsultation, København, 1952 (in Danish).

[6] D. Hoffmeyer, *Measurement of drum noise – A pilot project*, NT 1597-02, Nordtest, Finland, 2004

[7] NF S 31-074, *Acoustics – Measurement of sound insulation in buildings and of building elements – Laboratory measurement of in-room impact noise by floor covering put in this room*, Norme francaise, AFNOR, 2002

[8] EN ISO 140-8:1997*, Acoustics – Measurement of sound insulation in buildings and of building elements – Part 8: Laboratory measurements of the reduction of transmitted impact noise by floor coverings on a heavyweight standard floor*

[9] EN ISO 717-2:1996, *Acoustics – Rating of sound insulation in buildings and of building elements – Part 2: Impact sound insulation*

[10] U. Jekosch, J. Blauert: *A semiotic approach toward product sound quality*. Inter-Noise 96, GB-Liverpool, 1996. 2283–2286.

[11] A.-C. Johansson, E. Nilsson, *Measurement of drum sound*, NT 1636-03, Nordtest, Finland, 2004.

[12] E. Zwicker and H. Fastl, *Psychoacoustics, facts and models*, 2nd edn., Springer-Verlag, Berlin, 1999

[13] W. Fasold, Untersuchungen über den Verlauf der Sollkurve für den Trittschallschutz im Wohnungsbau, *ACUSTICA*, 1965, 15, 271-284.

[14] E. Zwicker and H. Fastl, *Examples for the use of loudness: Transmission loss and addition of noise sources*, Inter-Noise 86, Cambridge, Massachusetts, 1986, 861-866.

[15] E. Nilsson and P. Hammer, *Subjective Evaluation of Impact Sound Transmission Through Floor Structures*, Engineering Acoustics, LTH, Lund University, Lund, Sweden, TVBA 3103, 1999.

[16] E. Zwicker, Meaningful Noise Measurement and Effective Noise Reduction, Noise Control Engineering Journal, 1987, 29, 66-76.

[17] ISO/DIS 140-11*, Acoustics -- Measurement of sound insulation in buildings and of building elements -- Part 11: Laboratory measurements of the reduction of transmitted impact sound by floor coverings on lightweight reference floors*

[18] W. Scholl, Impact sound insulation: The standard tapping machine shall learn to walk!, *Building Acoustics*, 2001, 8(4), 245-256.

[19] D. Montgomery, *Design and analysis of experiments*, 5th edn., John Wiley & Sons, Inc., 2001.

[20] L. Cremer, M. Heckl, E. Ungar, *Structure-Borne Sound*, 2nd edn., Springer-Verlag Berlin, 1988.

[21] EN ISO 3744:1994*, Acoustics - Determination of sound power levels of noise sources using sound pressure – Engineering method in an essentially free field over a reflecting plane*

[22] P.V Rao and L.L Kupper, Ties in paired comparison experiments: A generalization of the Bradley-Terry model, *American State Associate Journal*, 1967, 62, 194-204, Corrigenda 63, 1550

[23] EPLF NORM 021029-2, *Laminate floor coverings – Determination of drum sound generated by means of a tapping machine*, EPLF, Association of European Producers of Laminate Flooring, Bielefeld, Germany, 2003

[24] ASTM E691 – 99, *Standard practice for conducting an interlaboratory study to determine the precision of a test method*

[25] EN ISO 140-2:1991, *Acoustics -- Measurement of sound insulation in buildings and of building elements -- Part 2: Determination, verification and application of precision data*