On Some Biases Encountered in Modern Audio Quality Listening Tests—A Review*

SŁAWOMIR ZIELIŃSKI, *AES Member,* **AND FRANCIS RUMSEY,** *AES Fellow*
(S.Zielinski@surrey.ac.uk) (f.rumsey@surrey.ac.uk)

Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, UK

AND

SØREN BECH, *AES Fellow*
(SBE@bang-olufsen.dk)

Bang & Olufsen, Struer, Denmark

A systematic review of typical biases encountered in modern audio quality listening tests is presented. The following three types of bias are discussed in more detail: bias due to affective judgments, response mapping bias, and interface bias. In addition, a potential bias due to perceptually nonlinear graphic scales is discussed. A number of recommendations aiming to reduce the aforementioned biases are provided, including an in-depth discussion of direct and indirect anchoring techniques.

0 INTRODUCTION

Formal listening tests are nowadays regarded as the most reliable method of audio quality assessment. For example, they are typically used for the evaluation of low bit-rate codecs [1], [2]. Although many standardized methods for the evaluation of audio quality have been established over the last 20 years (see [3] for a detailed review), three major recommendations emerged and are used most frequently. The first, standardized in ITU-R BS.1116 [4], was developed primarily for the evaluation of small impairments in audio quality. The second, commonly referred to as MUSHRA (ITU-R BS.1534-1 [5]), is intended for the evaluation of audio systems exhibiting intermediate levels of quality. The third method, as defined in ITU-T P.800 [6], is widely used for the evaluation of telephone speech quality. The test procedures involved in these three standards are the result of improvements undertaken by many researchers over the last two decades. Nevertheless there is still scope for further improvements since, as will be demonstrated in this paper, modern listening test methods are not bias-free.

The term "bias" is used in this paper primarily to describe systematic errors affecting the results of a listening test. However, a number of examples demonstrating pitfalls in experimental design, increasing the random error, will also be provided. Random errors are commonly observed in the results of listening tests as they manifest themselves by a scatter of scores for a given experimental condition. Among other experimental factors, they are predominantly caused by an inconsistency of individual listeners in the assessment of audio quality, and they may also originate from interlistener differences in the evaluation of audio quality. It is important to mention here that random errors are easy to recognize and easy to take care of. For example, if the scores for a given experimental condition are symmetrically distributed with a central tendency, the random errors can easily be averaged out by calculating the mean value.

In contrast, systematic errors are very difficult to identify as they manifest themselves by a repeatable and consistent shift in the data. Consequently they may go unnoticed by the researchers. Systematic biases are also difficult to get rid of by means of statistical postprocessing of data. Once a certain degree of systematic error has been introduced to the data, it cannot be averaged out by statistical methods [7]. Hence systematic errors can poten-

*Manuscript received 2007 August 17; revised 2008 April 8.

tially lead to misleading conclusions and, worse still, can propagate further if the data subsequently are used for other purposes. For example, according to Poulton, some degree of bias has already been perpetuated in the international standard for loudness [8]. If data with an unknown degree of systematic bias are used as a basis for the development of an objective algorithm for the prediction of audio quality, these biases may reduce the reliability of the development significantly. In addition it might be impossible to say to what extent the developed model predicts the genuine data and to what extent it predicts the biases in the calibration data. In view of these facts, awareness of the typical biases, the ways they manifest themselves, and their likely magnitude may be of help in the design of a listening test. Moreover, a knowledge of the biases may not only help to minimize their effect through careful experimental design but could also help with the interpretation of the data.

The purpose of this paper is to provide a systematic exposition of the key biases encountered in modern audio quality listening tests based on a literature review. A number of examples demonstrating different biases are provided. This paper is not intended to give an exhaustive discussion of all possible biases that could arise in listening tests but, as the title implies, to provide a review of “some” biases. The biases discussed in this paper were selected since they commonly affect the results of popular audio quality listening tests and many engineers are still not aware of them. The emphasis was placed on the bias due to affective judgments, the response mapping bias, and the interface bias (see Sections 3, 4, and 5, respectively). In addition a potential bias due to perceptually nonlinear graphic scales is discussed in detail (Section 4.7). The final part of the paper contains an in-depth discussion of the methods that can be used to reduce the biases described (Section 6). Table 1 summarizes the main biases discussed. It also presents a brief summary of possible manifestations of these biases and it gives some examples of how these biases can be reduced. References to the sections of this paper where these biases are described in detail are also included in the table.

There are many other biases encountered in the listening tests that, due to space limitation, are not discussed here. This includes bias due to audiovisual interactions and bias due to the evaluation of audio quality under divided attention. For a comprehensive review of the audiovisual interactions see [9]. Some preliminary reports demonstrating differences in the evaluation of audio quality under divided or undivided attention are published in [10], [11]. A general discussion of biases in quantifying judgments, presented from a psychological point of view, can be found in [8], [12]. In addition, an informative review of biases involved in sensory judgments, primarily in the food quality assessment, is presented in [13], [14]. A more specific overview of biases encountered in listening tests is included in [3], [15].

In order to introduce the uninitiated reader to the methodologies of listening tests and to provide a background for the paper, the first section reviews in brief the most common techniques used for the evaluation of audio quality. This paper is an extended and updated version of a paper presented in 2006 [16].

1 COMMON TEST METHODS

As was mentioned in the Introduction, there are three types of listening tests that are nowadays used most frequently. Their distinct features will be summarized briefly. The first common method, as standardized by ITU-R BS.1116-1 [4], involves a paradigm based on a triple-stimulus with hidden reference approach. During the test listeners have access to three stimuli at a time (A, B, and C) and can switch between them at their will. The first stimulus (A) is a signified reference recording explicitly labeled “reference.” The two remaining stimuli (B and C) consist of an unsignified reference, commonly referred to as “hidden reference,” and a processed recording. In every trial the hidden reference and the processed recording are assigned to stimuli B and C randomly. The listeners are normally asked to assess the basic audio quality of these two recordings in comparison with the reference recording. The basic audio quality is defined as the single, global attribute used to judge any and all detected differences between the reference and the evaluated recordings. The results of the evaluation are recorded by the listener on a graphic, continuous scale. An example of the assessment scale is presented in Fig. 1. Listeners are instructed to give a maximum score for the hidden reference. The scale used in this method is often referred to as an impairment scale. In this paper this scale will be referred to as ITU-R impairment scale, and the labels used along this scale will be called ITU-R impairment labels.

This method was originally developed for the evaluation of small impairments in audio quality. In order to evaluate audio systems exhibiting intermediate levels of audio quality more efficiently, a new method was developed [17], [18] and is now standardized in ITU-R BS.1534-1 [5]. This method is known under the acronym MUSHRA, since it involves a “multistimulus test with hidden reference and anchor.” In contrast to the previously discussed method, in a MUSHRA test the listener has access to many more stimuli, one of which is a signified reference and the remaining stimuli are under assessment. The exact number of stimuli may vary, but it is recommended that that no more than 15 items be included in any trial. Similarly to the previously described method, the listeners are normally asked to assess the basic audio quality defined as a single, global attribute representing any and all detected differences between the reference and the evaluated stimulus. In contrast to the previous method, in which the ITU-R impairment scale was employed, the listeners are asked to record their judgments on a so-called continuous quality scale, an example of which is depicted in Fig. 2. The scale is divided into five equal intervals representing five quality categories ranging from “bad” to “excellent,” as indicated by the labels. In this paper the scale used in the MUSHRA test will be referred to as the ITU-R quality scale, and the verbal descriptors associated

Table 1. Summary of main biases reviewed.

Bias Type	Manifestations	Potential Implications	Examples of Bias Reduction
Recency effect	Result of audio quality assessment is biased toward the quality of that part of an audio excerpt that was auditioned most recently (Sec. 2.1).	Over- or underestimation of audio quality.	Use short, looped recordings with consistent characteristics. Randomize temporal distribution of distortions or use a continuous method of audio quality assessment.
Bias due to equipment appearance, listener expectations, preference, and emotions	Systematic shift of scores. Bimodal or multimodal distribution of data (Secs. 3.1, 3.2).	Over- or underestimation of audio quality. Summarizing results by averaging scores across all listeners may be misleading.	Use blind listening tests. Use a large population of listeners with different backgrounds.
Stimulus spacing bias	Listeners tend to equalize the differences between the scores given to different stimuli regardless of actual differences between stimuli (Sec. 4.1).	Distorted information about genuine differences in quality between stimuli.	Use stimuli that are equally spaced in perceptual domain (impractical). Use systextual design (Sec. 6).
Stimulus frequency bias	Expansion effect in distribution of scores (Sec. 4.2).	Overestimated differences between most frequent stimuli.	Avoid presenting some stimuli more often than others.
Contraction bias	Compression effect in distribution of scores (Sec. 4.3).	Underestimated differences between stimuli.	Familiarize listeners with range of stimuli under assessment. Avoid monadic tests. Use a direct anchoring technique (Sec. 6.2).
Centering bias	Systematic shift of all scores (Sec. 4.4).	Impossible to assess absolute audio quality. Information about rank order is preserved.	Avoid multiple-stimulus tests. Use direct or indirect anchoring techniques. Use systextual design (Sec. 6).
Range equalizing bias	"Rubber ruler" effect. Scores span entire scale regardless of actual range of stimuli (Sec. 4.5).	Impossible to assess absolute audio quality. Information about rank order is preserved.	Avoid multiple-stimulus tests. Use direct or indirect anchoring techniques. Use systextual design (Sec. 6).
Bias due to perceptually nonlinear scale	Nonlinear effect in distribution of scores. For example, some differences between scores can be compressed, others expanded (Sec. 4.7).	Distorted information about genuine differences between stimuli. Information about rank order is preserved. Data should be treated as ordinal (use nonparametric statistical analysis).	Use only two labels at ends of scale with no labels in-between or use a label-free scale. Use technique of indirect scaling (Sec. 6.1).
Interface appearance bias	Quantization effect in distribution of data (Sec. 5).	Distorted information about genuine scores. Data should be treated as ordinal.	Avoid labels, numbers, or tick marks along scale. Use more than 30 listeners.

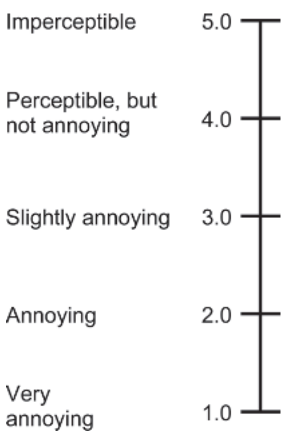

Fig. 1. ITU-R impairment scale [4], [19].

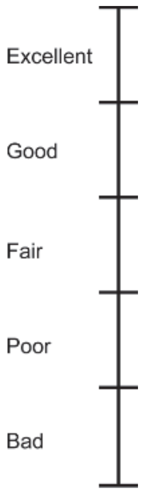

Fig. 2. ITU-R quality scale [5][19].

with this scale will be termed ITU-R quality labels [5], [19]. According to the MUSHRA standard, the hidden reference, being an unimpaired version of the original recording, has to be included as one of the items under assessment. In addition to the hidden reference, a mandatory low-quality anchor recording, namely, a 3.5-kHz low-pass filtered version of the original recording, has to be included in the pool of the items evaluated. Optional anchor recordings showing similar types of impairments as the systems under assessment can also be employed in the test. The listeners are instructed to assign a maximum score for the hidden reference; however, no instructions are given regarding how the listeners should assess the mandatory or optional anchors. The MUSHRA test is currently used widely to assess systems exhibiting intermediate levels of audio quality. In particular, this method is commonly used to evaluate the quality of low-bit-rate audio codecs [1], [2].

The third method that is in common use is standardized in ITU-T P.800 [6]. It is intended for speech quality evaluation. Although different variants of this method exist, in the most basic one (see [6, app. B]) listeners are sequentially exposed to phonetically balanced speech recordings (one recording at a time), and they are asked to evaluate the quality of each recording using five discrete categories, as illustrated in Fig. 3. In this method it is common practice to use a number of unsignified (unlabeled) anchors, called reference recordings, so that experiments made in different laboratories or at different times in the same laboratory can be sensibly compared [6].

There are many more methods in use. A comprehensive review of the current methodologies used for audio quality evaluation can be found in [3]. As mentioned, the three methods described in this section were included due to their popularity and to the fact that they can serve as good examples of the methods exhibiting biases that will be discussed later in this paper.

2 OUTLINE OF BIASES IN LISTENING TESTS

This section presents a brief overview of biases encountered in modern listening tests and forms a framework for more a detailed analysis of selected biases that is covered in the remaining part of the paper. In order to analyze the stages at which biases can occur, it might be useful to consider typical processes involved in the preparation of a listening test, its execution, and the analysis of the data obtained. These stages are depicted diagrammatically in Fig. 4.

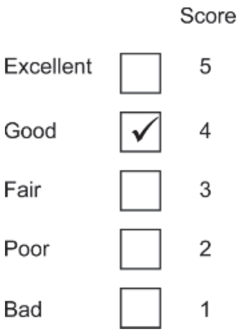

Fig. 3. Five-point category quality scale. (Adapted from [6].)

2.1 Selection of Stimuli

The selection of audio stimuli, as indicated in the top block in Fig. 4, is normally the first task undertaken by the experimenter during the preparation of the listening test. Even during this preparatory procedure significant biases may be introduced. There are typically three criteria applied in the selection of program material. Selected recordings should be 1) representative, 2) critical, and 3) consistent in terms of their characteristics. The first two

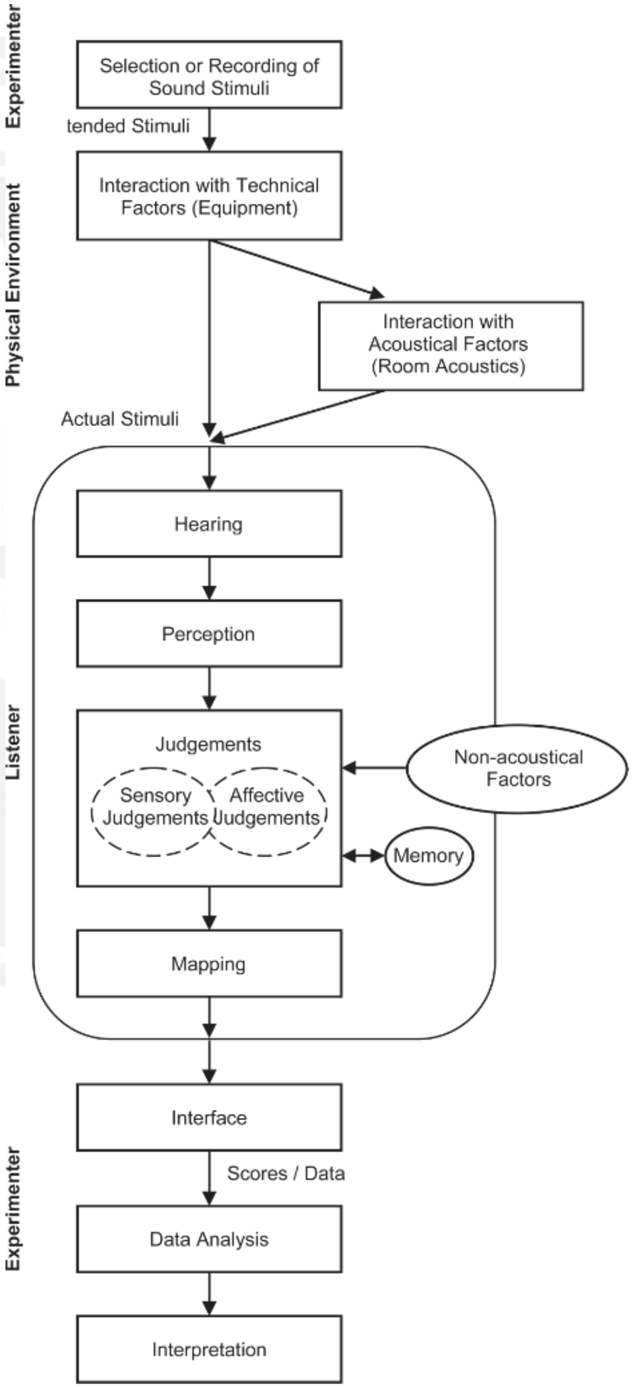

Fig. 4. Processes involved in preparation of listening tests, their execution, and analysis of results.

criteria, and their potential link to experimental bias, are discussed extensively by Toole [20], and also by Bech and Zacharov [3]. Therefore their analysis will be omitted here.

The third criterion, related to the selection of program material, is the consistency of characteristics. If the duration of a given excerpt is long, say more than 30 seconds, it is likely that its timbral and spatial characteristics will vary in time. If these variations are large, the listeners may find it difficult to "average" the quality over time, and consequently some random errors are likely to occur in the data (see [21] for an example). Therefore short, consistent, and perhaps looped excerpts are beneficial in this respect.

There is another problem related to using long, time-varying stimuli, which potentially can give rise to a systematic error. As mentioned, listeners face problems when evaluating the audio quality of long program material. It was observed that listeners are not reliable at "averaging" quality as it changes over the duration of the whole excerpt, and their judgments are biased toward the quality of that part of the recording that is auditioned last (the end of the recording if the recording is not looped). This psychological effect is related to the dominance of short-term memory over long-term memory and is often referred to as a recency effect, as the assessors tend to be biased toward recent events. For example, Gros et al. [22] conducted a study evaluating telephone speech quality and observed a systematic shift in scores due to the recency effect of a magnitude of up to 23% of the total range of the scale. Moreover, the recency effect was studied extensively by Aldridge et al. in the context of picture quality evaluation [23]. This phenomenon is sometimes referred to as a forgiveness effect as the assessors tend to "forgive" occasional imperfections in the quality, provided that the final part of the evaluated excerpt is unimpaired. For example, in the study conducted by Seferidis et al. [24] it was observed that for some stimuli the recency effect biased the results of the subjective evaluation by almost 50%.

There are three solutions to reduce the magnitude of the recency effect. The first, which is commonly used in audio listening tests, involves using short and consistent recordings in terms of their audio quality. The second solution is to randomize the temporal distribution of distortions for the same stimuli and use several profiles for the same levels of quality. Unfortunately this solution is expensive since it requires more stimuli and hence leads to a longer overall duration of the test. The third solution, which is sometimes employed in picture and multimedia quality evaluation experiments, involves a so-called continuous evaluation of quality. Instead of assessing the quality of a stimulus once, normally after its presentation, the participants are instructed to evaluate the quality of stimuli in a continuous manner during their presentation [25], [26].

2.2 Physical Environment

About thirty years ago a debate started in the audio engineering community as to whether the listening tests should be undertaken in a loosely controlled but more ecologically valid environment (such as a typical living room) or whether they should be undertaken in a highly controlled laboratory environment at a price of reduced ecological validity [27]. Nowadays we have strong evidence that the latter approach, although potentially leading to less ecologically valid results, has the advantage of higher sensitivity and accuracy over the former [28]. Consequently, in order to minimize the error and increase the sensitivity of the listening test, the physical environment has to be carefully controlled. ITU-R BS.1116-1 provides some of the most stringent guidelines in this respect [4]. More information about minimizing experimental errors arising from an inadequate physical environment can be found in [3, chapter on electroacoustic considerations].

2.3 Listener

The next four blocks in Fig. 4 (Hearing, Perception, Judgments, and Mapping) can be considered to form a holistic model of a listener as they describe physiological, psychological, and psycho-motor processes undertaken by the listener during the listening test. This part of the figure is similar to the so-called filter model of a listener described by Bech and Zacharov [3].

The Hearing block represents the psychological hearing properties of a listener. It is still unclear in what way and to what extent impairments in hearing properties may bias the results of the listening tests. Since the actual audio quality assessment is a high-level cognitive process, it might be possible that the brain can "compensate" for some small hearing impairments, especially when the test is performed at levels substantially exceeding the hearing threshold levels. Consequently it may be possible that the listener with small or even medium hearing impairments could perform similarly to a subject with normal hearing. This supposition is supported by the results of a study conducted by Bech [29]. However, according to Toole there is some evidence that the impairments in hearing threshold at frequencies below 1 kHz may lead to increased error in experimental data [15].

The subsequent block in Fig. 4, named Perception, represents those cognitive processes that allow the listener to describe the sound in terms of its basic characteristics such as loudness, timbral properties, pitch, and temporal and spatial properties. These and other nonaffective characteristics of the sound (such as sharpness, roughness) can be described, using Blauert's terminology, as sound "character" [30]. The ability to distinguish between different attributes of sound character varies across listeners. However, there is strong evidence that this ability can be developed by means of systematic training [31], [32].

The Perception block is followed by the Judgments block, which can be considered as the core component of a listener model in Fig. 4. As its name suggests, this block represents those judgmental processes that are responsible for the assessment of sound in terms of its character (sensory judgments) or in terms of its likeability (affective judgments). For example, according to sensory judgments, a given sound can be assessed as "quiet," "of high-pitch," "sharp," "bright," "coming from the front," "5 meters away," or "spanning 20 degrees of the frontal angle." In contrast to the sensory judgments, the affective judgments allow assessors to form and express their opinion as to how much a given sound is liked or disliked. For example, a given sound can be assessed as "pleasant," "nice," "likeable," "preferred," or "unpleasant," "disliked," "annoying," "irritating," and so on. An in-depth discussion of the bias in affective judgments will be presented in Section 3. The results of both sensory and affective judgments can be affected by memory, attention, and even non-acoustical factors such as visual cues. For example, Fastl reported that a train painted in red was perceived as being 20% louder than the same train painted in green [33]. Another example is reported by Guastavino et al., who observed that the fact that the listeners could see the subwoofer, even if it was not in use, affected their perception of the low-frequency audio content [34].

The last block in the listener model, Mapping, was included to represent the psycho-motor processes involved in the translation of internal judgments into the listener response. The exact form of the response is determined by the experimenter. Typically the listeners are instructed to express their internal judgments in terms of verbal categories or in terms of numbers. Recently graphic scales have been introduced to the listening tests, such as those presented in Figs. 1 and 2. The bias related to the mapping of scores will be discussed in more detail in Section 4.

2.4 Interface

Historically, paper-based questionnaires were used in order to enable an experimenter to collate the data. Nowadays computer-based interfaces are commonly used for this purpose. Depending on the nature of the experiment, they allow the listeners to record their judgments using a computer mouse and on-screen graphical objects such as tick boxes or sliders. The way the graphical user interface is designed may also have some bearing on the experimental error, which will be discussed in Section 5.

2.5 Data Analysis and Interpretation

The last two stages presented in Fig. 4 (Data Analysis and Interpretation) may also cause bias if they are handled improperly. Typical mistakes include choosing an inappropriate method of data analysis and omissions in checking whether the data meet statistical assumptions underlying the method. Due to space limitations these biases will not be discussed in this paper.

3 BIASES IN AFFECTIVE JUDGMENTS

As mentioned before, it is possible to distinguish between two types of judgments: sensory and affective. This give rise to a question as to whether the audio quality assessment involves only sensory judgments, or only affective judgments, or a combination of both. There are several arguments supporting the hypothesis that the evaluation of audio quality involves a combination of sensory and affective judgments. For example, in 1989 Letowski defined sound quality as "that assessment of auditory image in terms of which the listener can express satisfaction or dissatisfaction with that image" [35]. Eight years later, Blauert and Jekosch defined audio quality as "the adequacy of the sound attached to a product . . . with reference to the set of . . . desired features" [36]. These definitions refer not only to a sound character but also to affective terms such as dissatisfaction, adequacy, and desired features. In addition it is worth noting that verbal categories commonly used in the evaluation of speech quality and of intermediate levels of audio quality exhibit affective nature (excellent, good, fair, poor, and bad). The ITU-R impairment scale presented in Fig. 1 also contains an affective term (annoying). Hence it seems to be legitimate to conclude that the evaluation of audio quality does not involve only one type of judgments (sensory), but two types—sensory and affective—as was also pointed out by Västfjäll and Kleiner [37]. Consequently it is important to examine these two types of judgments in more detail. For the sake of a systematic description of different types of bias it is useful to separate a discussion on bias involved in affective judgments from bias involved in sensory judgments. Therefore in this section we will consider only biases involved in affective judgments. This will include a discussion on biases caused by the appearance of equipment, branding, situational context, personal preference, and even emotions and mood of the listeners.

3.1 Biases Due to Appearance, Branding, Expectation, and Personal Preference

In general listeners' affective judgments can be biased by the appearance of the equipment, price, and branding. For example, Toole and Olive demonstrated that in preference tests (affective judgments) both experienced and inexperienced listeners were biased by the appearance and the brand names of the loudspeakers evaluated [38]. When the scores from the listening tests were averaged across the listeners it was found that the results were different depending on whether the participants could see the loudspeakers or not. The maximum observed difference equaled 1.2 points (loudspeaker D, location 1) on a scale ranging from 0 to 10, which constitutes 12% of the range of the scale. This study is often quoted as a classical example of how important it is to undertake blind listening tests in order to reduce nonacoustic bias.

Listeners can also be biased by different labeling of the equipment. An interesting example is provided by Bentler et al. [39]. In their experiment a group of listeners were asked to assess the audio quality of two identical types of hearing aids, labeled as either digital or conventional. They found that out of 40 participants 33 listeners preferred the hearing aids labeled digital, 3 preferred the conventional ones, and only 4 participants did not hear the difference between the two. Like Toole and Olive, Benter et al. also emphasized the importance of undertaking blind listening tests in order to minimize this type of bias.

For a given object under evaluation assessors may give different scores depending on whether the object meets their expectations or not. It is likely that the participants will like the objects that meet their expectations and dislike any object that departs from their internal standard of expectation. For example, Fastl reported that the brand name of a car can trigger expectations about the sound character produced by a closing door [33]. Another interesting example is provided by Beidl and Stücklschwaiger [40]. They asked the listeners to do paired comparisons of different car noises. One group of listeners preferred quieter noises, whereas another group preferred louder noises, which was an unexpected outcome of the investigation. When asked for justification, the second group argued that "the higher the speed, the more powerful, the more sport(y), the more dynamic and, therefore, better." A substantial disparity of opinions between listeners was also observed by Rumsey in his listening tests investigating the benefits of upmixing algorithms [41]. According to his results, one group of listeners preferred original two-channel stereo recordings, whereas another group of listeners preferred upmixed five-channel recordings. Only a small proportion of listeners were "in the middle" with their neutral responses. The examples provided in the preceding indicate that in listening tests involving affective judgments, it is likely that the scores obtained for a given experimental condition will not be "polarized" toward one answer, but different opinions between listeners will emerge, giving rise to a bimodal or even multimodal distribution of scores. In this case a common practice of averaging the results across the listeners and summarizing the scores using some statistics, such as mean values and confidence intervals, may not only be illegitimate from a statistical point of view but, more importantly, could be misleading and may result in erroneous conclusions.

A more recent example of how the expectation of listeners may affect the results of an audio quality evaluation is given by Västfjäll [42]. In his experiment the expectation of the participants was controlled directly by asking them to read different consumer reports (either positive or negative). In the listening test participants were asked to evaluate the annoyance (affective judgment) of two different aircraft sounds. It was found that participants who had low expectations on average rated the unpleasant sound as less annoying than people who had high expectations. The difference was equal to about 13% of the total range of the scale.

An interesting example, demonstrating how nonacoustical factors, such as the meaning of the sound, can influence affective judgments is presented by Zimmer et al. [43]. They undertook a listening test investigating the unpleasantness of environmental sounds and then applied a probabilistic choice model based on physical characteristics of the stimuli in order to predict the data. Their model predicted the scores well for all sounds investigated except the "wasp" sound. They concluded that in this case the listeners' judgments of unpleasantness could have been governed by nonacoustical factors: "Based on the comments by participants some of whom reported to instinctively have ducked, or tried to wave off 'the bee from their left ear,' the excess annoyance of this sound may be tentatively characterized as being due to its intrusiveness." Another example demonstrating how the meaning of the sound may influence the affective judgments is provided by Fastl [33]. He reported that the bell sound may be interpreted by German subjects as "pleasant" and "safe," due to its association with the sound of a church bell. On the contrary, for Japanese listeners the sound of a bell may lead to feelings denoted by the terms "dangerous" or "unpleasant," since it could be associated with the sound of a fire engine or a railroad crossing.

Another problem related to the listening tests involving affective judgments is their poor long-term stability of results. Although in the experiments conducted by Choisel it was observed that the results of preference judgments were stable over a period of approximately six months [44], in general the listeners' preferences may drift over time due to, for instance, changes in fashion, changes in listening habits, or changes in the technical quality of available products. Consequently this may prevent researchers from drawing conclusions that would hold true for a long time. For example, in 1957 Kirk undertook a study investigating people's preferences in audio quality. According to his results, 210 college students preferred a narrow-band reproduction mode (limited to 90–9000 Hz) compared to the unrestricted frequency range [45]. Had this experiment been undertaken nowadays, it is likely that the results would have been different. Kirk also found that listeners' preferences can be changed by continued exposure to a particular audio system. The issue of the long-term stability of affective judgments in listening tests requires further studies.

3.2 Bias Due to Emotions and Mood

It is important to distinguish between emotion and mood. The former is a specific reaction to a stimulus, whereas the latter is a general "background" feeling. Both emotions and mood can have some effect on affective judgments, and there is some evidence that "happy" people make more positive judgments. For example, Västfjäll and Kleiner [37] investigated the effect of emotion on the perception of sound quality using the annoyance scale and found out that for some listeners the mood biased the results by as much as 40% with respect to the total range of the scale. Moreover, in a more recent experiment Västfjäll observed that listeners who had a positive frame of mind judged the pleasantness of sound significantly higher than people who had a negative attitude. In addition it was found that those listeners who were annoyed evaluated sounds higher on the annoyance scale compared to the listeners in a neutral mood [42]. The magnitude of this effect varied across listeners and ranged from 10% to almost 40% of the total range of the scale. These examples illustrate to what extent affective judgments of audio quality are prone to nonacoustic factors such as mood or emotional state. This is one of the reasons why it is advantageous to use many listeners. Using a large population of listeners, preferably at different times of the day, may help to average this bias out. If a listening panel contains a similar number of listeners with a positive frame of mind compared to those with a negative one, this bias may cancel out.

3.3 Situational Context Bias

Food scientists have observed that affective judgments may change depending on the situational context. For example, some food or beverage products are more liked in a restaurant than in a home setting. In other words, the same product may fit one situation and not another. This may imply that for a given sound stimulus, its quality may be evaluated differently depending on the situational context, and hence it might be advisable to conduct situation-oriented studies. For example, some levels of audio quality may be unacceptable in a carefully designed listening room but may be tolerable in a kitchen. Although some authors seem to support this hypothesis (see the definition of audio quality proposed by Blauert and Jekosch [36]), there are no direct data available in its support. On the contrary, there is some evidence contradicting this hypothesis. For example, Gros et al. [22] showed that the context of environmental or visual cues has a very weak influence on the audio quality of speech recordings. This is also in accordance with the findings of Beresford et al., who conducted a study investigating contextual effects of listening room and automotive environments on sound quality evaluation [46]. The result of their investigation showed that the listening context had no effect on scores when using the single judgment method. There was some evidence, however, that the null result was due to contraction bias, which will be discussed in Section 4.3. When the experiment was repeated with a multiple-stimulus method, some differences between the environments were found, but they proved to be very small [47].

4 RESPONSE MAPPING BIAS

As mentioned before, the task of the listeners in the listening tests is not only to judge the quality of sound but also to translate their internal judgments into some form of response, such as the position of a slider along the grading scale, which is symbolically expressed by the Mapping block in Fig. 4. The term “internal judgments” is used in this paper to denote the listeners’ implicit judgments that are made inside their minds. The mapping process of internal judgments onto external responses is not straightforward and, as will be shown, may involve some substantial biases.

It has to be acknowledged here that the distinction made in this paper between internal judgments and response mapping is to some extent artificial, and it might be possible that the listeners undertake these two tasks together as one task. There might therefore be a significant overlap between mapping and judgments. (The main difference is that the judgments, in contrast to mapping, are solely psychological processes and do not involve any motor processes.) The distinction between judgments and mapping was made in this paper for the sake of a systematic description of biases involved in the listening tests. However, one cannot exclude the possibility that the biases outlined in this section refer to mapping as well as judgments. In fact Poulton calls them “biases in quantifying judgments” [8]. He identifies, among others, the following biases: stimulus spacing bias, stimulus frequency bias, contraction bias, centering bias, and range equalizing bias. They will be briefly outlined next, followed by some examples found in the audio engineering literature.

4.1 Stimulus Spacing Bias

Fig. 5 illustrates a potential mapping error caused by stimulus spacing bias. The left-hand side of the figure shows a hypothetical distribution of judgments in the perceptual domain. This distribution can be considered as a genuine, bias-free distribution of internal judgments made by a listener for a given, hypothetical set of audio stimuli. The right-hand side of the figure shows the distribution of scores given by a listener in response to this hypothetical set of stimuli. Ideally, under bias-free conditions, both distributions should be identical. In other words, the scores on the right-hand side should be spaced proportionally compared to the unbiased judgments presented on the left-hand side of the figure. However, if the mapping process is affected by the stimulus spacing bias, the scores obtained in the listening test (the right-hand side of the figure) will be spaced at more or less equal intervals, regardless of the distribution of the stimuli in the perceptual domain. The arrows in Fig. 5 show how the internal judgments in the perceptual domain would be mapped onto the scores obtained in a listening test if the mapping process was affected by the stimulus spacing bias. It can be seen that small subjective differences between stimuli in the perceptual domain are expanded on the assessment scale, whereas large differences between stimuli in the perceptual domain are compressed on the assessment scale.

The stimulus spacing bias was investigated in more detail by Mellers and Birnbaum using visual stimuli [48]. It was also investigated by Zieliński et al. in the context of audio quality assessment [49]. They demonstrated that the

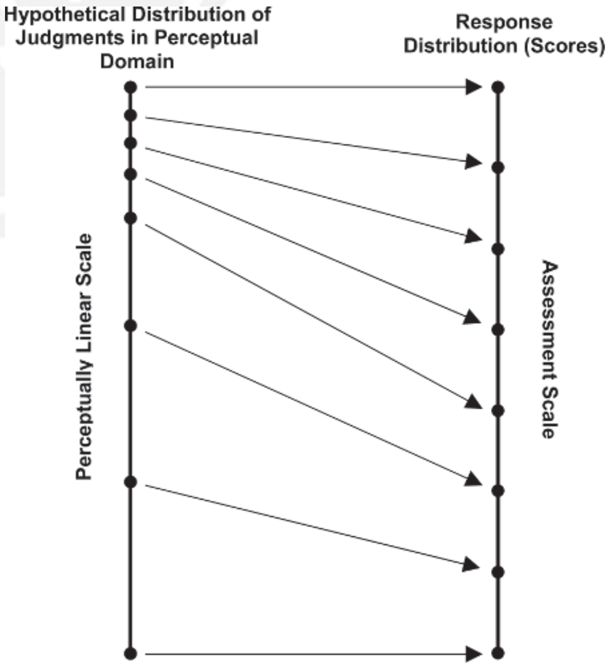
Hypothetical Distribution of Judgments in Perceptual Domain
Response Distribution (Scores)
Perceptually Linear Scale
Assessment Scale

Fig. 5. Stimulus spacing bias model. (Adapted from [8].)

results of the MUSHRA test can be affected by the stimulus spacing bias. The magnitude of the observed bias ranged up to approximately 20 points on the 100-point scale, which corresponds to a whole category change in terms of the quality labels used along the scale. For example, the quality of one of the recordings was assessed as "good" in one listening test and as "fair" in another. This discrepancy undermines the absolute meaning of the ITU-R quality labels. This issue will be discussed later.

4.2 Stimulus Frequency Bias

The stimulus frequency bias is illustrated in Fig. 6. As can be seen on the left-hand side of the figure, the second stimulus from the top is judged five times whereas the remaining stimuli are judged by the assessors only once. The stimulus frequency bias manifests itself by the fact that the observers treat the more frequent stimulus (or stimuli) as if they were nearly, but not exactly, the same. Consequently the responses to the most frequent stimuli occupy more of the response range than they should [8]. Under the bias-free condition one would expect that the second stimulus from the top should be assigned the same score every time it is presented to the assessors. However, as is demonstrated on the right-hand side of the figure, the scores assigned for this stimulus are scattered along the scale.

It can be argued that the stimulus frequency bias and the previously discussed stimulus spacing bias are of a similar nature as they manifest themselves in a similar way. The common feature of these two biases is that if the stimuli are not distributed equidistantly in the perceptual domain but there is a local maximum in the distribution, either due to an increased frequency of some stimuli or a large number of similar stimuli, the responses to the more frequent or similar stimuli will occupy a larger range of the assessment scale than they should (expansion effect).

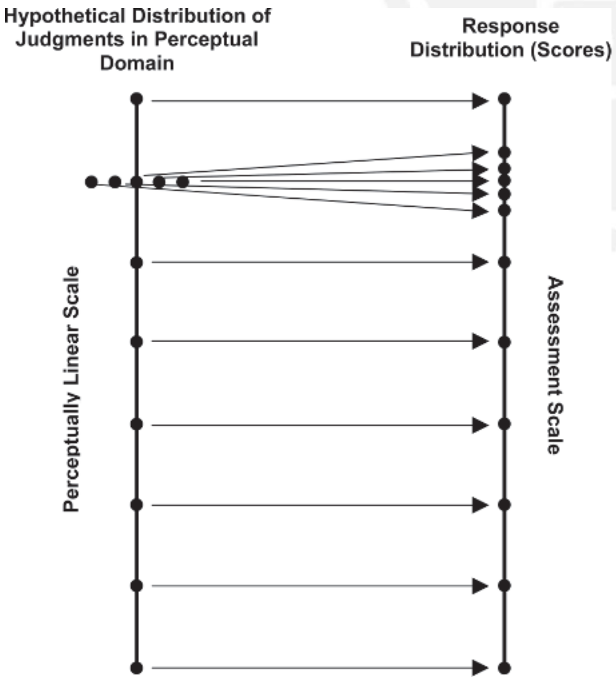
Hypothetical Distribution of Judgments in Perceptual Domain
Response Distribution (Scores)
Perceptually Linear Scale
Assessment Scale

Fig. 6. Stimulus frequency bias model. (Adapted from [8].)

The stimulus spacing bias and the frequency bias were studied by Parducci [50], [51]. He established a mathematical model describing these and other "contextual" biases. However, it is not clear whether his model can be used to correct the result of the MUSHRA test in order to account for the biases mentioned. This would require experimental verification.

Since the stimulus spacing bias and the stimulus frequency bias affect primarily the results obtained using multistimulus methods such as MUSHRA, a possible way of reducing this bias is to use the listening test method, in which every listener evaluates one and only one stimulus. This method is sometimes referred to as a monadic test. However, it will be shown in the next section that this solution can result in contraction bias.

4.3 Contraction Bias

Contraction bias can be regarded as a conservative tendency in using the grading scale as listeners normally avoid the extremes of the scale, which is illustrated in Fig. 7. Under the bias-free condition one would expect that the range of the scores on the right-hand side would be the same as the perceptual range of the stimuli presented on the left-hand side of the figure. Poulton [8] distinguishes between two subcategories of this bias: stimulus contraction bias and response contraction bias. In the case of the stimulus contraction bias the data tend to concentrate near the assessors' inner reference, or once the participants get familiar with the distribution of the stimuli, their judgments tend to be mapped around the center value of the distribution. In the case of response contraction bias, the results may be affected if the participants know the center value of the range of permissible responses (such as the midpoint of the scale). The center value becomes an "attractor" and listeners' responses tend to be biased toward it. It is worth noting that both subcategories of the contraction bias are often blended together, and they manifest themselves in a similar way. Therefore when the experi-

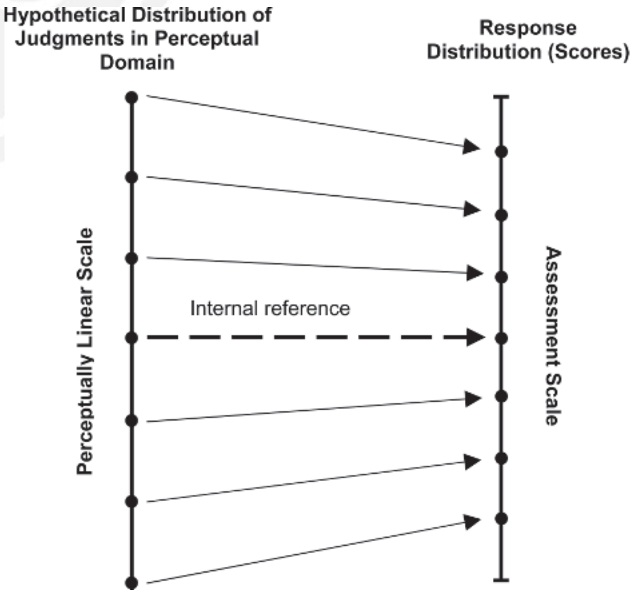
Hypothetical Distribution of Judgments in Perceptual Domain
Response Distribution (Scores)
Perceptually Linear Scale
Internal reference
Assessment Scale

Fig. 7. Contraction bias model. (Adapted from [8].)

mental data exhibit a contraction effect it may be difficult, without a detailed analysis of the experimental design, to ascertain to what extent the results have been affected by the stimulus and to what extent by the response contraction bias. Therefore because of their similarity we will treat these two subcategories of the contraction bias jointly, and no further distinction between them will be made.

The contraction bias primarily affects the results of the tests based on single-stimulus presentation, such as the method described in ITU-T P.800 [6]. In particular, its magnitude is the highest for a very first judgment, as the listeners exercise caution in using the scale at the beginning of the listening tests if they are not familiar with the stimuli.

An example of the contraction bias was observed in experiments undertaken by Beresford et al. [46]. In a series of different listening tests they investigated the quality of four identical stimuli in a car and in a listening room. They used three different listening test methods: 1) monadic test, 2) MUSHRA test, and 3) modified MUSHRA test, referred to as a multiple-comparison test. The audio stimuli were created by introducing different magnitudes of linear distortions (frequency ripples). They found that the results obtained in a car and a listening room were similar. However, they differed substantially depending on the test method used. Their results, extracted from trained listeners only, are presented in Fig. 8. In the monadic test each observer assessed one and only one stimulus during the experiment. For reasons of clarity, in the case of the monadic test, the 95% confidence intervals were omitted. They were relatively large and ranged up to ±28 points due to the difficulty in recruiting sufficiently large groups of trained listeners independently for each experimental condition. As can be seen, for the monadic test the mean results span only about 30% of the scale. The worst quality stimulus exhibiting 18-dB ripples was on average graded as 40, whereas the best recording (0 dB) was assessed as 73. This demonstrates the listeners' conservative tendency in using the scale and confirms that the contraction bias can be a problem if the test involves single judgments. In addition it is interesting to see that in the case of the monadic test the results obtained for the unprocessed stimulus (0 dB) and for the stimulus with 6-dB magnitude of spectral ripples are almost identical, which supports the view that listening tests that do not exploit any comparisons with other stimuli, which is the case of the monadic test, lack sensitivity, and hence possess poor resolution [27]. According to Lawless and Heymann, humans are very poor absolute measuring instruments; however, they are very good at comparing stimuli [13].

4.4 Centering Bias

Many audio engineers assume that modern listening tests allow for the assessment of audio quality in absolute terms. However, as will be demonstrated, the centering bias causes the scores to "float," rendering the assessment scale relative rather than absolute.

The centering bias is illustrated in Fig. 9. According to Poulton [8] it does not affect the relative distances between judgments; however, it determines the way in which the judgments are projected onto the grading scale. For example, Fig. 9 shows two hypothetical distributions of stimuli (sets A and B). If these two sets of stimuli are assessed independently, say by two different groups of listeners, in the extreme case of the centering bias the midpoints between the maximum and minimum values of the judgments for both sets will be projected onto the mid value of the scale. This effect can lead to severe errors in the results of the listening tests. For example, in the hypothetical model presented in Fig. 9, the judgments of stimuli *X* and *Y* (midpoints in the distributions on the left

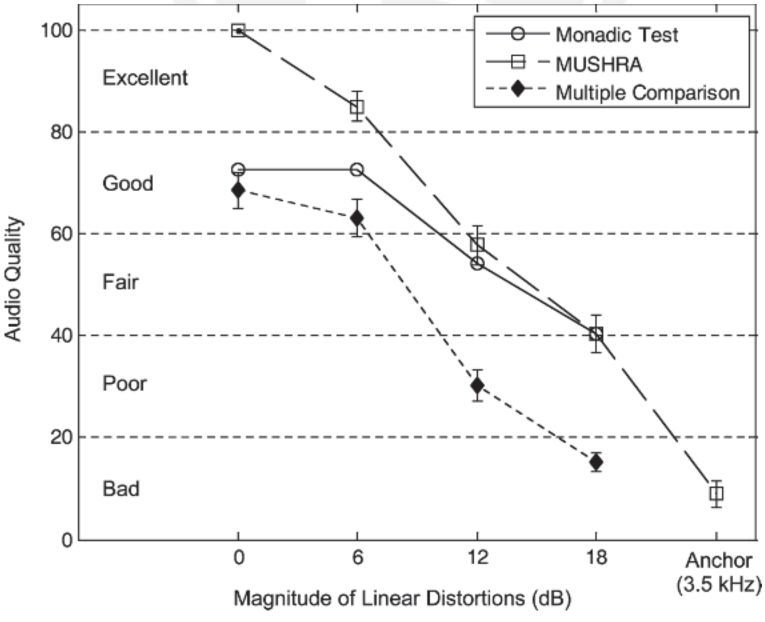

Fig. 8. Example of biases in listening test scores. (Results extracted from data obtained by Beresford et al. for trained listeners only [46], [47].)

and right, respectively), although perceptually accounting for different levels of audio quality, are erroneously mapped to the same scores. As can be seen, the absolute position of the judgments is shifted as the midpoints between the extremes are equalized.

An example of the phenomenon that could have been caused by the centering bias was reported by Toole [20]. After analyzing the results from a number of experiments, he identified a systematic shift in scores for some loudspeakers: “Some of the resident ‘reference’ loudspeakers have been seen to drift down the fidelity scale as superior products overtake them.” In another paper devoted to the evaluation of loudspeakers he reported an opposite effect, resulting in the upward shift in scores: “Given the opportunity to compare good products with inferior ones directly, listeners rated the good product higher than when the good products were compared with one another in isolation” [15]. The magnitude of the reported bias was rather small (approximately 3%).

The centering bias, among other biases, was studied systematically by Helson [52]. In his adaptation level theory he postulated that the judgment of a given stimulus does not only depend on its own properties but also on the properties of other stimuli. The result of a subjective evaluation is relative and tends to be biased toward the weighted mean of stimuli (typically logarithmic), called the adaptation level. For example, humans easily adapt to the ambient temperature, provided it is not too extreme, and become unaware of tactile stimulation from the clothing [13]. There is also some evidence that people adapt to loudness [53]. In addition, Toole states that we adapt to the room acoustics [54]. It is worth noting that Helson’s “adaptation” is a very broad term, encompassing not only the centering bias but also other biases described in this paper, such as the contraction bias and the stimulus spacing bias. It also includes so-called sensory adaptation, which can be defined as a decrease or change in sensitivity to a given stimulus as a result of continued exposure to that stimulus or a similar one [14]. Due to the broad scope of this term it might be useful to distinguish between physiological and psychological adaptation. The former term could refer to physiological changes in peripheral organs due to a change in the intensity of stimuli, whereas the latter term could describe psychological biases, such as the centering bias described in this section. The centering bias, due to its nature, is also sometimes referred to as central tendency [55], [56].

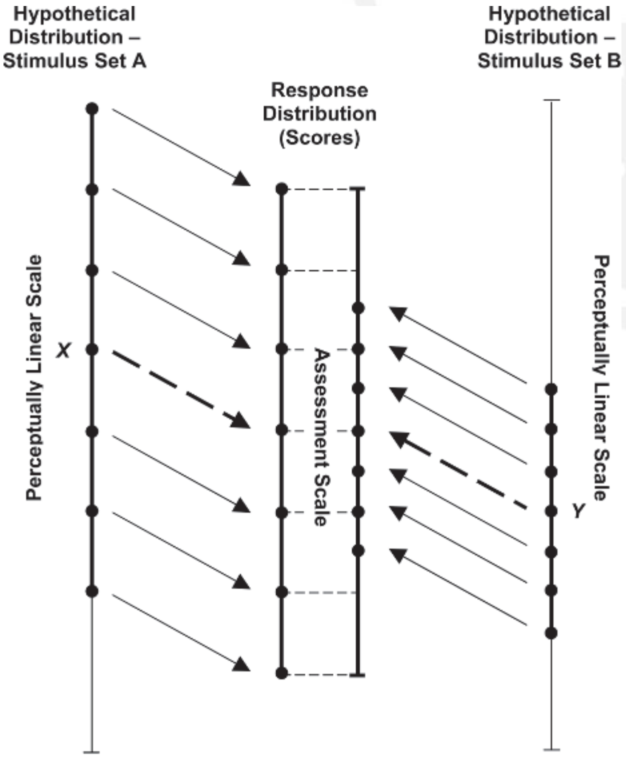
Hypothetical Distribution – Stimulus Set A
Response Distribution (Scores)
Hypothetical Distribution – Stimulus Set B
Perceptually Linear Scale
X
Assessment Scale
Perceptually Linear Scale
Y

Fig. 9. Centering bias model. (Adapted from [8].)

As shown in the graphic model in Fig. 9, the centering bias introduces a systematic error in the projection of genuine judgments onto the grading scale, manifesting itself in the erroneous shift of scores on the scale. This gives rise to a question as to whether it is at all possible to design a listening test eliciting the absolute quality levels of audio stimuli. One may argue that due to the problems with the centering bias and the range equalizing bias, which will be discussed next, it might be difficult, if not impossible, to achieve the stated goal. This supposition is confirmed by the results obtained by Bramsløw [57]. In his experiment he attempted to design a listening test evaluating the absolute values of selected characteristics of sound such as clearness, sharpness, spaciousness, and loudness. His listening panel consisted of both normal and hard-of-hearing listeners. The results were surprising as the average values of the scores obtained from the normal and from the hard-of-hearing listeners were the same in the statistical sense. This means that, contrary to expectation, the scores from the two groups of listeners overlapped. Their mean values were centered at the same level on the scale. According to Bramsløw, given the simplistic hearing-aid algorithm provided for the hard-of-hearing listeners, it is unlikely that the two groups of listeners had the same auditory perception of the stimuli. Nevertheless, “they rated the mean values equal, giving a strong indication that the judgments obtained on the subjective scales were not absolute” [57]. In the opinion of these authors it is likely that the effect of equalization of the midvalues of the scores obtained from the two groups of listeners was caused by the centering bias. This demonstrates the difficulty in the design of a listening test aiming at the assessment of absolute levels of audio quality.

4.5 Range Equalizing Bias

Another problematic bias that makes it difficult to assess the absolute levels of audio quality is the so-called range equalizing bias, and its model is presented in Fig. 10. A hypothetical bias-free distribution of judgments corresponding to two different sets of stimuli (A and B) is shown on the left- and the right-hand sides of the figure, respectively. It is assumed that the stimulus sets A and B are judged independently, for example, by two different groups of listeners. As can be seen, regardless of the range of the distributions of the stimuli, the scores in both cases span the whole range of the assessment scale. In other words, in the stimulus range equalizing bias the assessors use the same range of responses regardless of the size of the range of audio quality levels. It is not surprising, therefore, that Lawless and Heymann described this effect in terms of a “rubber ruler” since the assessors tend to “stretch” or “compress” the grading scale (ruler) so that it encompasses any stimulus range [13]. As a result the scores span the whole scale, regardless of the range of the stimuli. The range equalizing bias often occurs in multiple-stimulus methods, such as the MUSHRA test. This has been recently confirmed experimentally by Zieliński et al. [49]. The magnitude of the range equalizing bias depends on the number of stimuli used in the test since, according to Parducci’s range frequency theory, the size of the (contextual) effect increases as a function of the number of stimulus levels [50], [51].

The problem with the range equalizing bias, and with the other biases discussed in this section, is that they normally do not occur in isolation. When the data from the listening tests are analyzed it is difficult, or sometimes impossible, to identify precisely which biases affected the scores, as they can occur simultaneously. For a given experimental set, some of the biases may exhibit a higher magnitude than others. Sometimes it is even possible that some biases will cancel each other. The graphical models presented here can be considered as exaggerated examples of the biases discussed so far. Normally the observed effects are of a much smaller magnitude. For these reasons many authors exercise caution with the precise identification of the biases observed in the experimental data, and they tend to use generic but ambiguous terms such as contextual effects [48], [50], [58] or the already discussed term of adaptation effect [52].

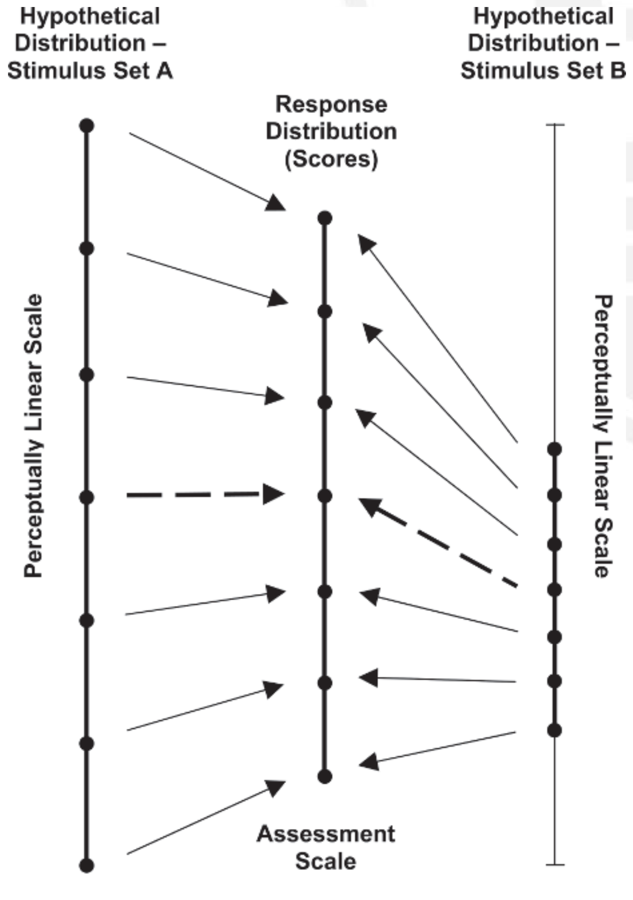
Hypothetical Distribution – Stimulus Set A
Hypothetical Distribution – Stimulus Set B
Response Distribution (Scores)
Perceptually Linear Scale
Perceptually Linear Scale
Assessment Scale

Fig. 10. Range equalizing bias model. (Adapted from [8].)

4.6 Further Examples

Despite these difficulties in the precise identification of biases affecting the data, a number of examples of the biases in the mapping of scores can be found in the literature. In ITU-R BT.1082-1 [7], for instance, an interesting example of a bias is described. When a group of assessors were asked to judge the picture quality of “high-quality” unimpaired HDTV recordings, they used all the terms from the quality scale, including “poor” and “bad.” When they were questioned regarding “bad” or “poor” pictures, they said that there had been none. Nevertheless, during the subjective test they used all available quality categories. This may indicate that the assessors did not evaluate the quality in an absolute sense but projected their internal judgments on the whole range of available quality categories without reference to the meaning of the labels, which could be explained by means of the range equalizing bias. In addition, this example shows that the assessors did not use the quality labels in the absolute sense of their meaning, which may indicate the lack of (absolute) meaning of the ITU-R quality labels in subjective tests.

Interesting examples of the range equalizing, centering, and contraction biases can be found in Fig. 8. As was mentioned, the data coming from the monadic test exhibited a strong contraction bias as the scores are grouped within the midpart of the scale (solid line). For the multiple-comparison test the scores span a higher range of the scale and are located predominantly in the mid and bottom parts of the scale. It is interesting to see that for the MUSHRA test the results are almost identical to the results obtained in the multiple-comparison test, but they are shifted in parallel to the top of the scale. The reason for the shift of the reference (0-dB stimulus) to the top of the scale is the direct anchoring technique used in MUSHRA (see Section 6.2). In contrast to the monadic and the multiple-comparison tests, an extra stimulus was used in the MUSHRA test. It was a 3.5-kHz anchor. As can be seen in Fig. 8, the anchor was evaluated using the scores at the bottom of the scale, giving the average score of 9. Since the bottom of the scale was occupied by the scores obtained for the above anchor, it can be speculated that the scores obtained for the 18-dB stimulus, which in the multiple-stimulus method was graded at the bottom of the scale too (rubber ruler effect), had to be “pushed up.” This effect can be attributed to the range equalizing bias. This upward shift in scores can also be explained by the centering bias model. For example, it can be argued that due to the inclusion of the low-quality anchor in the MUSHRA test, the midpoint of the distribution was lowered. Hence the judgments for all stimuli had to be projected upward on the scale, and consequently the 18-dB stimulus was graded higher than it should be. In addition it is interesting to notice that the magnitude of the bias demonstrated in

Fig. 8 reached up to 30% of the total range of the scale and hence changed the meaning of the results based on the labels. For example, the quality of the 6-dB stimulus was “good” according to the multiple-comparison test and “excellent” according to the MUSHRA test results. Similarly, the quality of the 12-dB stimulus fluctuated between “poor” and “fair,” depending on the test method. This indicates that the listeners did not regard the meaning of the labels in their absolute semantic sense. Hence it might be concluded that the quality labels are obsolete and that the listening tests could be undertaken with the scales without the labels. This supposition requires further experimental verification.

4.7 Bias Due to Perceptually Nonlinear Scale

In a typical listening test it is easy to prove statistically that some audio processes under test are better than others. However, without being sure that the scale used in the experiment was perceptually linear, it might be impossible to say how much the processes differ in a perceptual sense. Hence it might be problematic for broadcasters or product designers to make certain economic decisions if they do not know their likely impact in terms of the magnitude of a change in the perceived audio quality.

Up to this point in our discussion we implicitly assumed that the assessment scales used in the listening tests were perceptually linear. However, this may not be the case, and if the scales used in the listening tests are not perceptually linear, it is likely that a systematic error will affect the results of the test. This error may prevent the researchers from reaching correct conclusions about the perceptual differences between the evaluated stimuli. The graphical model of this potential bias is presented in Fig. 11. The right-hand side of this figure shows an example of an assessment scale with the standard ITU-R quality labels commonly used for the quality evaluation. Typical listening tests are designed in such a way that the assessment scales are numerically linear and the verbal labels attached to the scales are equidistant, as it is illustrated in Fig. 11. Consequently the distance between adjacent labels is the same in the numerical sense. For example, in speech quality evaluation scores ranging from 1 to 5 are commonly assigned to the quality labels [6], as is illustrated in Fig. 3. The same scores (1 to 5) are also assigned to the five equidistantly distributed labels used in the ITU-R impairment scale (see Fig. 1). Another example of a numerically equidistant interval scale is the ITU-R quality scale used in the MUSHRA test (see Fig. 2). Although in this case the labels are not anchored to particular points but to the five intervals uniformly distributed along the scale, the numerical distances between the adjacent intervals are also equal.

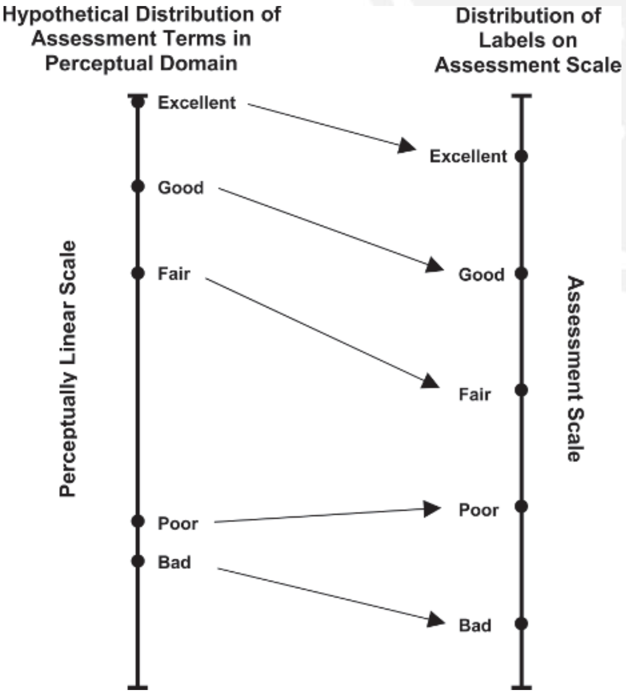
Hypothetical Distribution of Assessment Terms in Perceptual Domain
Distribution of Labels on Assessment Scale
Perceptually Linear Scale
Assessment Scale
Excellent
Good
Fair
Poor
Bad
Excellent
Good
Fair
Poor
Bad

Fig. 11. Model of bias due to perceptually nonlinear distribution of labels.

Because of the equidistant distribution of labels and because of the numerically linear properties of these scales, some researchers implicitly assume that the perceptual distances between adjacent labels are also equal. If this assumption is not checked, it may be difficult, or even impossible, to make unbiased perceptual inferences based on the listening test results. For example, if in a given experiment exploiting a 100-point scale an experimenter observed a 20-point upward shift of scores, both at the bottom and at the top of the scale, it might be possible to reach a conclusion that the same perceptual improvement of quality was observed in the case of low-quality and high-quality items. However, this conclusion may be incorrect if the scale is perceptually nonlinear as, for example, a 20-point shift of scores at the top of the scale could perceptually account for much greater quality improvement than a 20-point shift at the bottom part of the scale.

The left-hand side of Fig. 11 shows a hypothetical distribution of the standard quality assessment terms in the perceptual domain plotted on the perceptually linear scale. Although the distribution presented here is referred to as hypothetical, it is in fact based on the empirical results of semantic scaling of a group of adjectives, which was part of the study conducted by Watson [26]. She found that British English speakers regard the terms “poor” and “bad” as similar, but the terms “poor” and “fair” as distinctly different, which is reflected in the nonequidistant distribution of the quality terms on the perceptually linear scale. As can be seen in Fig. 11, if the nonequidistant label points on the perceptual scale are mapped onto the equidistant label points on the assessment scale, the perceptual distance between “fair” and “poor” will have to be compressed, whereas the perceptual distance between “poor” and “bad” will have to be expanded, giving rise to the bias of perceptually nonlinear mapping of judgments.

A potential risk of this bias was a concern of many researchers working in the field of picture and multimedia quality evaluation. For example, Narita undertook a study in Japan [59] using the Japanese translation of the ITU-R quality and impairment labels. In contrast to Watson’s findings discussed before, he observed that the semantic differences between adjacent labels were approximately the same. Likewise the results of a study conducted in Germany using the German equivalents of the ITU-R quality and impairment labels showed that the differences between adjacent terms were approximately the same [7]. However, the results of similar studies conducted in several other languages, such as French [7], Italian [60], American English [60], Swedish [61], and Dutch [62], revealed substantial semantic differences between adjacent labels, which are illustrated in Fig. 12. For clarity reasons this figure contains only the mean values, and any error bars were omitted. However, the degree of variation in the available data was examined and compared between the studies (results not presented here). The 95% confidence intervals ranged from 4% (with respect to the whole range of the scale) for the study conducted in the Dutch language to 12% for the study conducted in the Swedish language. The data regarding the variations of scores in the studies conducted in France and Germany were not available. As can be seen in Fig. 12, the biggest semantic differences between adjacent labels of the ITU-R quality scale were observed in England in the study undertaken by Watson [26] using a group of native British English assessors. Likewise, a substantial variation in the differences between adjacent labels was also observed in the experiment involving the American English assessors across different states of the U.S. [60]. As can be seen in Fig. 12, for British English (UK) the terms “excellent,” “good,” and “fair” are located at the top of the scale, whereas the terms “poor” and “bad” are clustered together in the lower part of the scale. As mentioned, this creates a substantial gap between the labels “fair” and “poor.” These differences indicate a potential risk of perceptual nonlinearity of the studied scales. It is argued that the semantic differences across different linguistic groups, cultural groups, and between individuals can cause a misrepresentation of the results by as much as ±50% [7].

In addition to the standard ITU-R quality labels it was also decided to present in Fig. 12 the available data for the universally used term OK, as it is an interesting example of semantic variations between languages. (These data were not available in German and Dutch.) For instance, the term OK has a quite positive meaning in Italian, equivalent to “buono;” however, it has a rather neutral meaning in American English and Swedish, semantically similar to “fair.” Although the results presented in Fig. 12 show semantic differences between the ITU-R quality labels, similar but smaller effects of nonequidistant spacing were also observed in the case of the ITU-R impairment labels, such as those used in ITU-R BS.1116 [4].

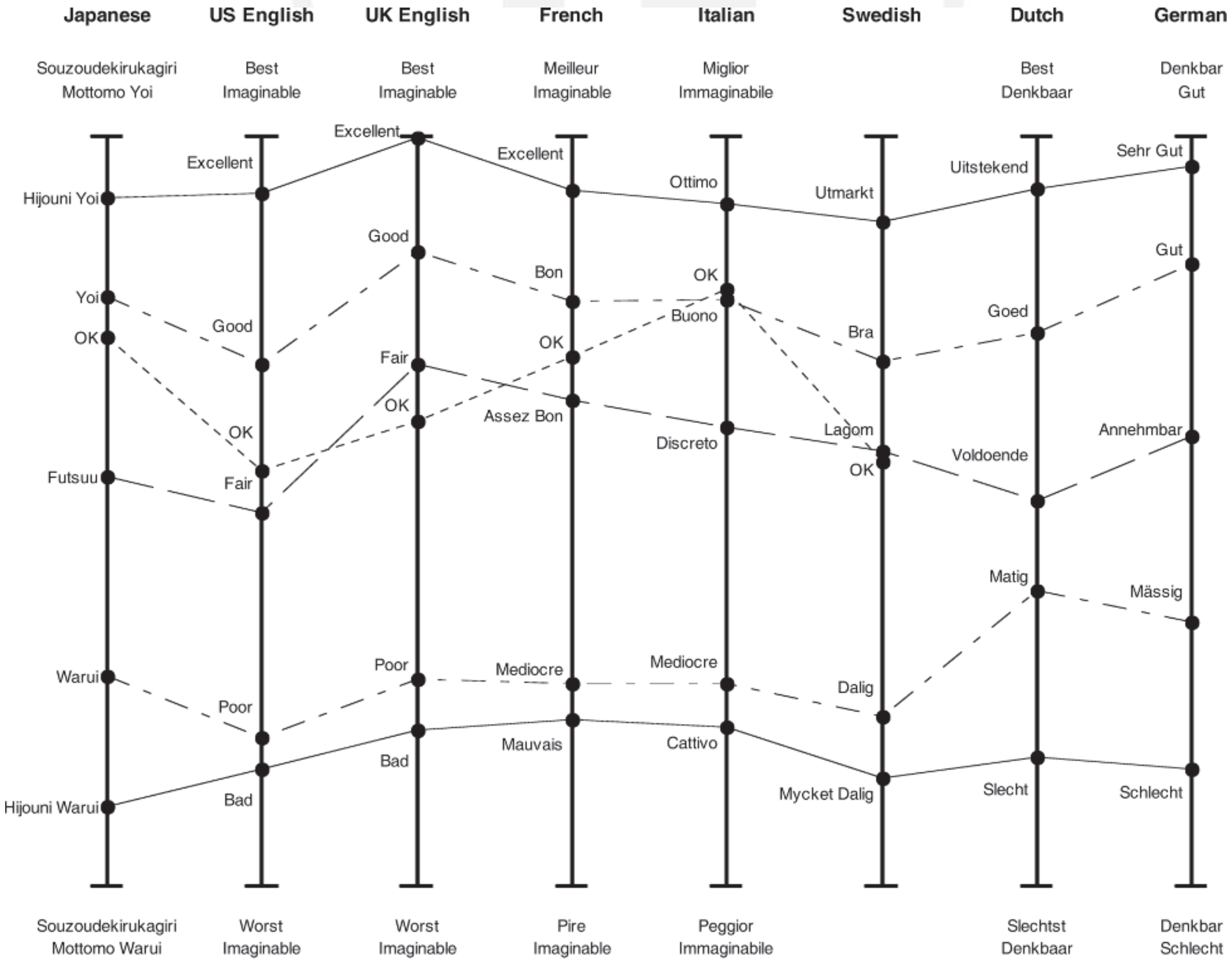

Fig. 12. Combined results of scaling of labels used in quality evaluation. (Data taken from [7], [26], [59]–[62]; see text for details.)

Considering a large variation in semantic differences between adjacent quality labels for most of the languages and assuming that the model of a bias due to a perceptually nonlinear scale presented in Fig. 11 is correct, it is possible to conclude that the ITU-R quality and impairment scales are perceptually nonlinear, with the exception of the German and Japanese equivalents of these scales. For this reason some researchers, such as Virtanen et al. [61], questioned the equal-interval property of the ITU-R quality scale and concluded that it shows "profound nonlinearity." Watson also criticized the ITU-R scale and stated that it is invalid and should not be used as an equal-interval category scale. She expressed her concern about the fact that this scale is so popular: "That it continues to be used all over the world by telecommunications companies, in face of evidence that it is not a reliable method, is alarming at best" [26]. She also argued that in order to circumvent the problem of the nonlinearity the labels on the scale should be removed. She proposed an unlabeled scale called "polar scale," consisting of a vertical line 20 mm long with no labels other than a + sign at the top and a − at the bottom to indicate the polarity of the scale. In fact, this proposal is very similar to Note 1 in ITU-R BS.1116. According to this note, the use of predefined anchor points (labels) may introduce bias. As an alternative it is recommended to use "the number scales without description of anchor points" with an indication of the intended orientation of the scale [4]. A similar solution is proposed in ITU-R 1082 and ITU-T P.910, which recommend using graphic scales with only two labels at the extremes [7], [63]. A scale with the labels at the endpoints and no labels in between was also recommended by Guski [64]. Despite these recommendations, many researchers still employ scales with labels.

As was concluded before, scales without labels could be used instead of labeled scales if one wanted to avoid a problem with nonlinearity of a scale. In order to verify this approach, Zieliński et al. [65] recently undertook an experiment in which they compared the results obtained using the labeled ITU-R quality and impairment scales with the results obtained using the unlabeled scale for an identical set of audio stimuli and the same listening paradigm. Since according to the literature summarized here the labeled scales exhibit nonlinear properties, it was expected that the results obtained using the labeled scales would be substantially different from the results obtained using the unlabeled scale. Contrary to what was expected, when the results obtained using these scales were compared against the results obtained using the label-free scale, the differences between them were negligibly small. It is difficult at this stage to provide a reliable explanation for why the listening tests yielded almost the same results in all three cases. However, one possible explanation is that the listeners ignored the meaning of verbal descriptors along the scale and used the graphic scale without reference to the labels, or perhaps only taking the endpoint labels into account. If this supposition is correct and if it is confirmed by future experiments, the use of labels may be rendered obsolete, and consequently it might be advisable to undertake listening tests using label-free graphic scales.

5 INTERFACE BIAS

It will be shown in this section that the design of the interface for the listening tests is an important issue as it may lead to an extra experimental bias.

Early listening tests involved a form of tape-based playback, which posed some restrictions on both the experimenter and the listeners. For example, the participants normally had to listen to one stimulus at a time. They had to complete the test at a fixed pace, determined by the playback system. Moreover it was difficult for the experimenter to fully randomize the order of stimuli for every listener, and hence it was more difficult to counter any learning effects [8] due to a fixed order of presentation (randomization of stimuli may help to average out these effects). The listeners typically were recording their judgments using a pen and paper.

One of the major advancements in the design of the interfaces used in the listening tests is the development of switching devices allowing the users to switch "instantaneously" between the stimuli and to compare them directly, which made it possible to design double- or even multiple-stimulus listening tests. Originally these devices were implemented using analog hardware [66], but recently they were replaced by computer-based systems. In order to make the switching less disruptive (seamless), different versions of compared recordings are often looped, and the switching between the samples is undertaken synchronously with about 40 ms cross-fade to avoid potential clicks [4]. The main advantage of being able to compare the audio stimuli directly is the benefit of using short-term memory. As was mentioned, this allows the user to detect small differences between the stimuli [27], [28]. In addition, a benefit of multiple-stimulus tests is the reduction of contraction bias, as discussed in Sections 4.3 and 4.6. Another advantage of switching devices is that they allow the users to undertake the test at their own pace. If necessary, the listeners can spend more time listening to more "challenging" items, which has a potential of reducing the random error.

The advantage of the computer-based interfaces is their interactivity. However, there are also some problems related to them. For example, some of the interfaces, especially those used for multiple-stimulus tests, can get quite complicated as they involve many graphical objects. As a result there is a scope for user mistakes. This issue was recognized in the latest version of the MUSHRA standard, where it is recommended to disable all on-screen objects that do not correspond to the currently auditioned stimulus in order to prevent the listener from making mistakes in using the interface [5]. This may decrease the experimental error.

The way the graphical user interface is designed, in particular the layout of the assessment scale, may have some bearings on the experimental data. Verbal labels attached to the scale, numbers, or even ticks on the scale may introduce distortions in the distribution of the data. Listeners seem to use the points of the scale that are associated with labels, numbers, or ticks more frequently than the remaining parts of the scale. Consequently the scores in the experimental data are clustered near the points where the labels, numbers, or ticks were attached. An example demonstrating how a visual appearance of the interface can bias the results is presented in Fig. 13 (data taken from [67]). This interface was developed for a test in which the listeners were asked to assess a certain attribute of four recordings (stimuli 1–4) using vertical sliders. In order to calibrate the scale, a direct anchoring approach was taken, with stimuli A and B determining the anchor points on the scale. The assessment scale is presented on the left of the figure between the A and B buttons and the first slider. As can be seen, the scale contained five major tick marks: two at the ends of the scale, two near the A and B buttons determining the anchor points, and one in the middle. In addition eight minor tick points were added to the scale. In order to help listeners to make visually easier comparisons between the relative positions of the four sliders, both major and minor tick marks were extended using the horizontal lines spanning the whole interface. This resulted in distortions in the distribution of the data. As it can be seen in the histogram presented on the right-hand side of Fig. 13, the scores obtained in the listening test are clustered near the horizontal lines. This quantization effect manifests itself by distinct peaks in the histogram, indicated by the asterisks. It can be seen that the scores are not distributed uniformly along the scale but exhibit a high degree of quantization effect near the horizontal lines. Another example demonstrating similar effects can be found in [68]. In order to avoid a problem of the quantization effect in the data, Guilford suggests that the scale should have no breaks or divisions [55].

It is difficult to assess how detrimental this effect is. However, from the statistical point of view, if the number of listeners taking part in the test is large, say greater than 30, it is likely that when the mean values across listeners are calculated, the quantization effect in the distribution of scores will be averaged out due to the central limit effect, and consequently this error may be neglected. According to the central limit theorem, for a large number of observations the distribution of the means tends to be normal (bell shaped), regardless of the distribution of the raw data [69]. Nevertheless, as a measure of precaution it might be advisable to inspect the distribution of the data in each experimental condition and the distribution of all data to check the magnitude of this error. If the number of listeners is small, say less than 10, it is likely that the quantization effect will not be averaged out, which may lead to errors. In that case if the distribution of the data for each experimental condition is not normal due to the quantization effect and if the variance between the conditions varies, the data should be treated as ordinal. Consequently it might be questionable to use some parametric methods of data analysis such as the analysis of variance. Under this circumstance the experimenter may be forced to use less powerful nonparametric techniques.

Another issue related to the interface design is its ergonomics and ease of use. For example, in MUSHRA-conformant tests some researchers tend to include in the interface 13 recordings for evaluation, or even more. For this number of stimuli there are 78 possible paired comparisons. Although this number of comparisons may seem to be manageable within one experimental session, it might be difficult for the listener to remember all of them in order to make accurate relative judgments. For example, after making 10 different comparisons it may be difficult for the listener to remember what the perceptual difference between the first compared pair of stimuli was, and hence the subsequent judgments may be less accurate. From a psychological point of view there are some limits on the amount of information that listeners are able to receive, process, and remember. There is some evidence that seven could be an optimum number of stimuli in this respect [70]. Another solution improving the ergonomics of the interface, and hence potentially reducing the assessment error, consists in allowing the listeners to sort the recordings evaluated according to the current positions of the sliders. For example, in the MUSHRA test the listeners could initially rank the recordings according to their quality and then sort them. As a result the buttons and the corresponding sliders would be rearranged according to the previously determined rank order. This procedure could be repeated many times in an iterative way, each time improving the accuracy of the assessment. The ad-

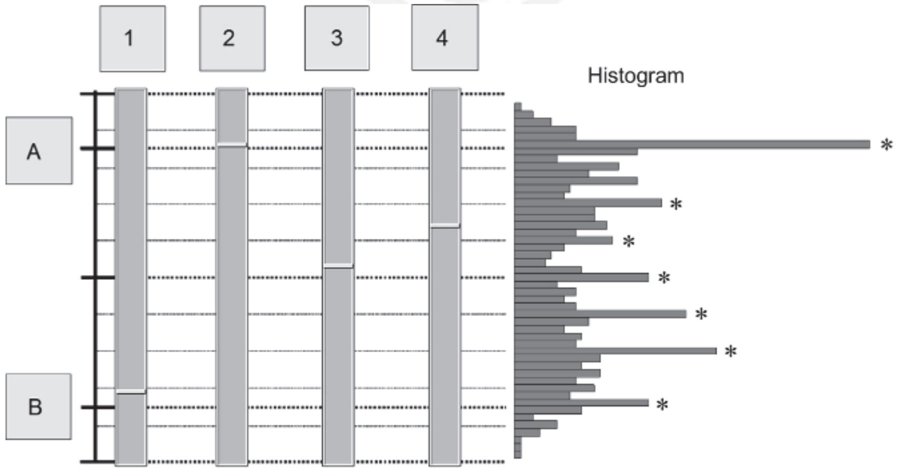

Fig. 13. Quantization effect in distribution of data due to horizontal lines crossing sliders. (Data taken from [67].)

vantage of this solution is that the listeners can focus their attention on comparisons of the adjacent items rather than comparing all of the items in the interface. One of the first examples of the use of this technique in the MUSHRA test can be found in [71].

6 REDUCING BIAS

In this section we will discuss the ways in which the biases can be reduced. It will be shown that some of the proposed methods may be expensive, as they require extra experimental effort, whereas others may introduce extra biases as a side effect. A brief summary of the main biases discussed in this paper as well as the methods of reducing their effects are presented in Table 1.

In Section 3 different biases inherent to affective judgments were discussed. Although this issue is still debated, considering the magnitude of the biases, potential problems with the distribution of data, and the risk of a drift of results in time, it might be advisable to avoid the tests involving affective judgments, if possible. This suggestion is in line with Köster [72], who claims that the problems related to affective judgments cast "serious doubts on the predictive validity of hedonic and consumer studies that rely on single measurement sessions." However, this solution may not always be feasible, since affective tests are often required by decision-making bodies as they provide useful information about market behavior. Under these circumstances it might be advisable either to undertake a number of listening tests, say with different groups of listeners, in order to increase the validity of the results, or to undertake the affective test in parallel with the sensory test, as the latter can provide additional descriptive information about the evaluated products.

Although more experimental evidence would be required to conclude which biases are biggest—sensory or affective—it is probably safe to assume that these biases are different and therefore may lead to different errors within one scale, if the scale is inconsistent and requires affective judgments at one end and sensory ones at the other. The ITU-R impairment scale can serve here as an example. The top of this scale can be considered sensory, as it is labeled "imperceptible," whereas the mid and bottom parts of the scale involve both sensory ("perceptible") and affective judgments ("annoying") [16]. Including both sensory and affective labels on the same scale may be confusing for the listener. A possible solution to this problem was suggested by Nielsen [73], who proposed to replace the current ITU-R impairment labels with the following terms: "not audible," "just noticeable," "small," "medium," and "large." Although the linearity of such a scale may be debatable, its advantage is that all the labels are only of a descriptive (sensory) nature. Another solution would be to use the scale proposed by Meilgaard et al. [14], which comprises the following terms: "imperceptible," "slightly perceptible," "moderately perceptible," "strongly perceptible," and "extremely perceptible."

Another advantage of the labels proposed by Nielsen and Meilgaard et al. is that they may help to extend the operational range of the scale as they do not contain the term "annoying." ITU-R BS.1116 was originally intended for the assessment of small impairments only. However, it is known that when this method is used for the evaluation of small impairments in audio quality, the listeners use only the top part of the ITU-R impairment scale (perhaps to avoid the remaining 75% range of the scale that corresponds to the term "annoying"), which reduces the operational range of the scale considerably [73].

Section 4 discussed the biases in the mapping of scores, including stimulus spacing bias and stimulus frequency bias. There are no easy methods of reducing the stimulus spacing and frequency biases. One possible way to reduce these biases is to make sure that the quality distribution of the audio recordings in each listening session is uniform. Poulton describes an iterative procedure that can be used to create such a set of stimuli [8]. However, this procedure is likely to be impractical in most audio engineering applications as it requires flexibility in adjusting the quality levels of the stimuli used in the test. In real-life applications the experimenters may not be able to adjust the quality levels of the evaluated recordings. For example, in the case of listening tests investigating the quality of low-bit-rate codecs the experimenter has only limited control over the distribution of the quality levels exhibited by the codecs under test.

There are three ways in which the contraction bias can be reduced. The first, suggested in [8], [3], [5], is to familiarize the listeners with the stimuli prior to the listening test. The familiarization procedure allows listeners to learn the range of the stimuli and, more importantly, it helps them to establish a rule by which they will map their judgments. The second method is to use a multiple-stimulus test, as opposed to a monadic test or single-stimulus test, examples of which were discussed in Section 4.6 (MUSHRA and multiple-comparison tests). The third way in which the contraction bias can be reduced, or even removed, is to apply the technique of direct anchoring (see Section 6.2).

Two biases that are probably the most difficult to reduce are the centering bias and the range equalizing bias, as they are responsible for a substantial shift of the scores on the scale and for a "rubber ruler" effect. As was discussed earlier, it is primarily these two biases that make it difficult to assess the audio quality in absolute, rather than relative, terms. Unfortunately these biases cannot be neutralized by a randomization of stimuli or counterbalancing [13]. According to Poulton [8], these biases can be reduced by designing tests involving only single judgments (monadic tests). However, there are two major problems with this solution. First, it is expensive and impractical as it requires a large number of listeners in order to make the results statistically significant. Second this solution is not bias-free either. As was discussed in Section 4.3, monadic tests are often affected by a strong contraction bias. Since there is no easy solution to this problem, instead of attempting to avoid these two biases it might be more advisable to control them using direct or indirect anchoring techniques. If these two biases are properly controlled, they can even be exploited to the advantage of the experiment. The main drawback of this solution is the loss of the capability to assess the absolute levels of the audio quality. However, considering that most researchers are interested in exploring the relative differences between audio stimuli rather than their absolute magnitudes, this method might be considered a useful alternative compared to the solution discussed previously. Due to their importance, both direct and indirect anchoring techniques will be discussed separately in more detail in Sections 6.2 and 6.3.

The idea of designing the experiments in which biases, such as the centering bias or range equalizing bias, are not avoided but carefully controlled was extended further by Birnbaum [12]. He proposed so-called systextual design, which involves a systematic change of the experimental context in terms of the stimuli. In systextual design the range and distribution of stimuli are manipulated systematically, and their influence on the results is analyzed. By applying this approach to the experimental protocol more information about the results can be found, including information about the magnitude of the biases and their resilience to the variations in the range of the stimuli or their distribution. Hence the conclusions can be reached with more confidence, and consequently more generalizable theories governing the results may be established. The drawback of this method is its high cost as it requires extra listening tests.

In Section 4.7 a bias due to a perceptually nonlinear scale was discussed. Although this problem was studied by many researchers over the last 20 years, no major solution has emerged yet, although at least four different methods have been proposed. For example, one of the solutions, suggested by Jones and McManus [60], is to employ a ratio scale. The ratio scale was originally introduced by Stevens [74] and has been used extensively in basic psychophysics for more than 50 years. The main advantage of the ratio scale is a lack of the "ceiling" effect as the scale is open-ended. In this method the listeners do not use any verbal labels but only numbers. Typically a number 1 is assigned to a certain reference stimulus. This can correspond, for example, to the audio quality of a traditional stereo system. If a system under evaluation exhibits two or three times better quality than the standard one, the listeners are instructed to assign to it a grade of 2 or 3. The participants are free to use any positive number, and hence they are not limited in terms of the upper end of the scale as the scale is open-ended. This feature of the ratio scale is of particular importance if new high-quality products are evaluated in the subjective tests, since the assessors are not limited by the upper end of the scale if they want to use much higher grades than they used to use in the case of traditional systems. However, research has shown that the ratio scale is not bias-free either [8], [75]. For example, Narens [76] provided a mathematical proof that Stevens' ratio-scaling method is invalid. This might be a possible reason why the ratio scale has not been introduced to any of the audio or picture quality evaluation standards.

The second possible solution to the problem of the perceptually nonlinear scale is to design a graphic scale with the labels uniformly distributed in a perceptual sense. This could be done easily on the basis of the available data of semantic differences between different adjectives presented in Fig. 12. The idea of "adjusting" the position of the labels on the scale in order to make the intervals between adjacent labels perceptually equal was originally suggested by Virtanen et al. However, they decided not to pursue it. They argued that "the descriptors may represent diverse semantic fields, which make them noninterchangeable" [61].

The third possible solution to the problem of the perceptually nonlinear scale is to use a scale with only two labels at its ends. This approach is in accordance with Note 1 in ITU-R BS.1116 [4]. Unfortunately not many researchers used this method for audio quality assessment and consequently, due to its small popularity, it is difficult to assess its validity.

Finally, the last (fourth) method that could be used to overcome the problem of the perceptually nonlinear scale is based on the indirect scaling technique. Due to the importance of this approach and the fact that this technique can be used to reduce other biases discussed in this paper, this method will be discussed in more detail next.

6.1 Indirect Scaling

In this method the listeners are asked to compare one pair of stimuli at a time and to answer a question regarding their similarity or difference. The question that the listeners are asked is usually very simple as it requires only binary answers, such as yes/no, louder/quieter, or better/worse. There are two distinct advantages to this approach. First, unlike graphic scales or category rating scales, it is considered to be the easiest method in terms of its use, as the participants only need to answer a simple question regarding the perceptual difference between two stimuli. Second, a number of rigorous statistical models exist allowing the researchers not only to test the validity of the method but also to convert the data of the paired-comparison tests into data represented on a continuous scale, such as the ratio scale [77]. Therefore this method is often referred to as an indirect method of scaling, as the data on the ratio or interval scales are obtained indirectly by means of a test involving the paired-comparison approach. For example, Zimmer and Ellermeier applied this method for the assessment of unpleasantness of sound [43], [78]. This method has also been applied recently by Choisel and Wickelmaier for scaling auditory attributes of reproduced multichannel sound [44], [79]. The major drawback of this method is its high cost. Since the statistical methods used to convert the data are based on probabilistic models, it is important that the pool of the stimuli under test exhibit different levels of perceptual similarity between the stimuli in order to obtain a range of probabilities of choosing one stimulus over another. For example, if the pool of items contained only good and bad recordings, without any intermediate quality levels, it is likely that the probability of assessing one type of recording as better than the other would be only 100% or 0%. Consequently, without any intermediate probability levels it would be impossible to use probabilistic models to convert the data obtained in the listening tests into scores presented on a continuous scale. Therefore it is important that the pool of items intended for assessment exhibit a range of quality levels, which normally requires inclusion of a large number of stimuli. Since this approach is based on paired comparisons and considering that this method normally requires a large number of listeners, the overall time required from all listeners for the assessment of all stimuli under test might be long, and hence the test might be time consuming. Hence this method may not be practical for the industry, where the listening tests often have to be undertaken quickly due to a short time scale of the development process. However, it might be a suitable approach for academic or regulatory organizations, where the validity and accuracy is of prime importance.

6.2 Direct Anchoring

The term “anchor” is used in an ambiguous way in the literature. It is often used to describe a verbal term, which defines a certain point on a graphic scale. For example, the labels on the scales presented in Fig. 1 could be described as anchors (or verbal anchors, to be more specific). However, the term “anchor” can also be used with regard to a stimulus defining a certain point of the scale. In this case it is not the word but the perceptual impression evoked by an anchor stimulus that defines a certain point on the scale. In this paper anchor is used according to the latter meaning of the term.

Guilford distinguished between two types of anchors. The first type represents the anchors that are not subject to evaluation (he calls them background anchors). The second type represents the anchors that are included within the pool of evaluated items and are also evaluated [55]. Consequently it might be possible to refer to them as foreground anchors. The latter category of foreground anchors can be further divided into two categories known as direct and indirect anchors [80]. In the case of the direct anchoring technique, the anchors are used to define certain points on the grading scale. The anchor stimuli are often signified in the interface as reference recordings. In contrast to direct anchoring, in the indirect anchoring technique the listeners are not informed about the inclusion of the anchor recordings and hence no instructions are given with their respect.

In the simplest case of direct anchoring, two stimuli are used to determine some characteristic points on the scale. Typically they are the maximum and minimum stimuli, and they are normally used to determine the endpoints of the scale. In this way the anchors help to calibrate the scale as they set a permanent “yardstick” [55]. In a more sophisticated version of this technique, extra stimuli can be used to anchor intermediate points on the scale. The technique of direct anchoring is often used in the sensory evaluation of food in order to calibrate the scale. For example, the spectrum descriptive analysis method [14] provides a description of a set of ingredients and recipes for the preparation of food stimuli anchored to specific grades on a scale.

The direct anchoring technique can be used as an effective means of controlling the range equalizing bias. If the anchors defining the ends of the scale are selected in such a way that they encompass the range of the stimuli in all listening sessions, the range of the stimuli will be fixed and consequently the range equalizing bias will be constant. This has a potential of stabilizing the results. This scenario can be considered as a deliberate and controlled exploitation of the range equalizing bias, since the stimuli, regardless of their range and their absolute perceptual magnitude, will always be projected on the whole scale.

The technique of direct anchoring is to some extent already used in the listening tests. For example, in the experiments designed according to ITU-R BS.1116 and MUSHRA, the listeners are instructed to evaluate the original recording (reference) at the top of the scale [4], [5]. The bottom of the scale is not directly anchored.

There are four potential advantages of direct anchoring: 1) reduction of the contraction bias, 2) improved resolution of the test, 3) calibration of the scale, and 4) stabilization of the results (higher degree of repeatability). These advantages will be briefly discussed next.

A proper use of direct anchoring, especially if it is used at both ends of the scale, may remove the contraction bias completely by increasing the span of the scores to the whole range of the scale. This, in turn, may increase the resolution of the test, which can be considered a big advantage of this technique. In addition this technique can also help to reduce the centering bias [8].

Another advantage of this method is the possibility of calibrating the scale. For example, if the recordings having known perceptual properties are used to anchor the ends of the scale, the perceptual frame of reference is defined precisely as the listeners are required to assess the stimuli in terms of the perceptual differences between anchors and evaluated stimuli [75]. Hence the scale can be considered calibrated with respect to the anchors used and therefore can be referred to as a “standard” scale using Guilford’s terminology [55]. Since the frame of reference is defined precisely, it might be easier for the experimenter to interpret the results obtained in the listening tests. Moreover a precise calibration of the scale may be of particular importance if the data from the listening test are intended to be used for the development of objective methods for the prediction of listening test scores. In this case it is essential that the frame of reference for the evaluation of the audio quality is defined precisely and that the biases, such as the range equalizing bias, are either reduced to a minimum or kept the same, regardless of the range and distribution of the stimuli under evaluation. If the scale is not properly calibrated, any potential biases in the data could propagate and may adversely affect the prediction capability of the objective methods for audio quality assessment.

Finally, it is known that both direct and indirect anchoring techniques can help to design experiments that will achieve a high degree of repeatability as the range equalizing bias is kept constant. This leads to more stable results. For example, Marston and Mason [2] recently undertook a large-scale MUSHRA-based experiment evaluating the effects of the codec cascading on the audio quality. The experiment involved five listening tests executed in five different countries. Considering the risk of a potential problem of perceptually nonlinear scale discussed in Section 4.7, the results obtained for the 3.5-kHz anchor and for the 10-kHz anchor were stable, with the differences being less than 10% with respect to the total range of the scale. This demonstrates that the technique described may help to achieve a high repeatability in the experimental procedure.

The disadvantage of the technique of direct anchoring is that due to the range equalizing bias any information about the absolute levels of the audio quality of the evaluated recordings is lost. Hence this method is not suitable for experiments aiming at an investigation of the absolute levels of audio quality. However, considering that the rationale for undertaking many listening tests is often to explore relative differences rather than absolute quality levels, this disadvantage may be considered irrelevant.

In the example of direct anchoring presented, the anchor recordings determined the endpoints of the scale. There is a slight risk, however, that some of the items evaluated may be either “better” than the top anchor or “worse” than the bottom anchor. Under these circumstances the information about these items would be lost due to the end-of-scale effect. In order to counter this problem, Guilford suggested that anchors should be set at a little distance from the ends of the scale in order to allow room for expansion if needed [55]. An example of this type of anchoring is presented in Fig. 13. As can be seen, anchor recordings A and B were used to define two points near the ends of the scale. A similar approach is used in some of the food-quality assessment standards, where it is recommended to use a 150-mm-long scale with anchors located approximately 15 mm from each end [56].

6.3 Indirect Anchoring

In the case of indirect anchoring the listeners are not made aware that any anchor recordings are introduced in the experiment and consequently are expected to evaluate them in a similar way as other items under evaluation. An example of the indirect anchoring technique can be found in the MUSHRA standard [5], where one mandatory and a number of optional anchors are recommended. As was mentioned, the first, mandatory anchor is a 3.5-kHz low-pass filtered version of the original recording. The optional anchors can include the low-pass filtered versions of the original recording with cutoff frequencies of 7 and 10 kHz, respectively. They may also include other types of impaired signals, depending on the nature of the test. The indirect anchoring technique is also commonly used in speech-quality evaluation. For example, according to ITU-T P.800 it is standard practice that a range of anchor recordings obtained by introducing different levels of noise to the original recording are included in the test [6].

There are several advantages to the indirect anchoring technique. The first is that, similarly to the direct anchoring technique, it has a potential to reduce the centering bias. If the anchors are selected in such a way that they encompass the range of quality levels exhibited by the items under test, it will be the anchor recordings that will determine the midpoint of the range of all the items evaluated in a given experiment, not the stimuli under test. Since the centering bias depends on the midpoint of the range of evaluated stimuli, as long as the anchor recordings encompass the range of the stimuli under evaluation, the midpoint of the range of all stimuli will be constant, and consequently variations in the scores due to the centering bias will be kept to a minimum. This may have a stabilizing effect on the experimental results (the results may be more repeatable). As mentioned, the results obtained by Marston and Mason for the anchor recordings were very similar in five different laboratories [2], which demonstrates the potential of the anchoring technique in stabilizing the results. It has to be stressed here that indirect anchors may help to stabilize the results provided that the anchors encompass the range of the stimuli to be assessed; in other words, they are maximum and minimum stimuli in terms of quality.

The second potential advantage of the indirect anchoring technique is that it may also help to calibrate the scale, although it is not as effective as the previously discussed direct anchoring technique. For example, the anchor recordings can be selected in such a way that they perceptually encompass all the other stimuli (they effectively become maximum and minimum stimuli in the perceptual sense). Now if the total number of evaluated stimuli is large, according to the Parducci range–frequency theory [50], [51], it is likely that the scores will span almost the entire range of the scale, and as a result one anchor recording will be assessed using the scores from the top range of the scale, whereas the second anchor will be evaluated using some scores from the bottom range of the scale. As a result this effect is similar to the effect of direct anchoring described, and hence may be exploited to calibrate the scale, although there is no guarantee that this technique will be as efficient as that of direct anchoring.

The third and probably biggest advantage of the indirect anchoring technique is that it provides an effective diagnostic tool that can help detect the presence of biases affecting the results. As was mentioned, some experimental biases are unlikely to be removed entirely, even in a carefully designed experiment. Hence it is vital that some means of checking for their possible presence and their likely magnitude be incorporated in the listening test. This task can be achieved with the help of the indirect anchoring technique. The results obtained for the anchors in different experiments, or even in different listening sessions of the same experiment, can be compared. If they are the same, it might be possible to conclude either that the biases are negligibly small (unlikely) or that they manifest themselves in the same, repeatable way. However, if the results are different, it could be an indication of bias. In this case there may be a need for further experimentation in order to explore the reasons for the observed variations in the data.

One of the first examples demonstrating the use of the indirect anchors for diagnostic purposes was provided by Sperschneider in 2000 [81]. In his experiment the listening sessions were blocked according to different bit rates of the audio codecs. When he examined the results obtained from different listening sessions, he noticed a small but systematic shift of the scores obtained for the anchors (up to 9%). A similar tendency was also observed by Wüstenhagen et al. [18]. In both cases it was concluded that this shift could have been caused by a change in quality of the evaluated items, since the scores obtained for the anchors dropped when the quality of the evaluated items was higher. This could be considered a classic example of the range equalizing bias. However, to some extent this effect could also be attributed to other biases, such as the centering and the stimulus spacing bias. Further experimentation would be required to investigate this phenomenon in more detail (systextual design).

7 SUMMARY AND CONCLUSIONS

This paper gives a systematic review of typical biases encountered in modern audio quality listening tests, with particular emphasis on the following three types of biases: bias due to affective judgments, bias encountered in the mapping of scores, and interface bias. The main biases discussed are summarized in Table 1.

It was shown that bias due to affective judgments may result in errors of up to 40% with respect to the total range of the scale. These errors can be caused by factors such as personal preferences, appearance of the equipment, expectations, and even mood of the listeners. In addition it was shown that bias due to the mapping of scores may also result in a similar range of errors. Considering the large magnitude of the bias in the mapping of scores and the fact that the results strongly depend on the actual range and distribution of the stimuli, it was concluded that the results obtained in the listening tests exhibit relative properties. For the same reason it was also concluded that it might be difficult, if not impossible, to design a listening test that aims to elicit absolute levels of audio quality.

Some examples shown in this paper, especially those based on the MUSHRA test, demonstrated that listeners' scores can be biased by up to 20% of the range of the scale, which corresponds to a change by one category in terms of the labels used along the scale. This may indicate that the absolute meaning of the labels used in the ITU-R quality scale is questionable. If this conclusion is confirmed by further research, labels used along graphic scales might be rendered obsolete, and the standard labeled scales could be replaced by label-free scales.

Considering the relative nature of the results obtained in most listening tests, in particular in the MUSHRA test, and evidence that the listeners do not always make absolute references to the verbal terms along the scale, a common practice of interpreting and presenting the results using references to the labels may be incorrect.

An issue of the perceptual properties of the graphic scales commonly used in the listening tests was also discussed. According to the literature, the verbal descriptors attached to the ITU-R graphic scales render them nonlinear. However, the most recent experimental evidence contradicts the conclusions reached in the past and shows that the departure from linearity is much less severe than was indicated by previous research on this topic.

The bias related to the user interface was also discussed. It was shown that the visual appearance of the scale and associated graphical objects may lead to a severe quantization effect in the distribution of scores. This may introduce errors to the results and may prevent the researchers from using parametric statistical techniques for data analysis. Some recommendations aiming to improve the functionality of interfaces were made.

The final part of the paper provides a description of several methods that can be used to reduce the discussed biases. The direct anchoring technique is of particular importance as it has a potential to calibrate the scale using precisely defined sound stimuli. Because of this property the technique of direct anchoring might be particularly useful if the data from the listening test are intended to be used for the development of objective methods for predicting listening test scores.

It is hoped that this paper will raise the awareness of researchers about the biases encountered in modern audio quality listening tests and will help to reduce some of them. Moreover, the examples discussed here may also help experimenters to identify the presence of potential biases in their data, which could enhance the analysis and interpretation of their results. Finally it has to be acknowledged here that some issues discussed in this paper are still debatable as, for example, the recommendation to use label-free scales. It is therefore hoped that these issues will prompt other experimenters to undertake further research in this area.

8 REFERENCES

[1] G. Stoll and F. Kozamernik, "EBU Listening Tests on Internet Audio Codecs," Tech. Rev. 283, European Broadcasting Union, Geneva, Switzerland (2000).

[2] D. Marston and A. Mason, "Cascaded Audio Coding," Tech. Rev. 304, European Broadcasting Union, Geneva, Switzerland (2005).

[3] S. Bech and N. Zacharov, *Perceptual Audio Evaluation. Theory, Method and Application* (J. Wiley, Chichester, UK, 2006).

[4] ITU-R BS.1116-1, "Methods for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunications Union, Geneva, Switzerland (1994).

[5] ITU-R BS.1534-1, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," International Telecommunications Union, Geneva, Switzerland (2003).

[6] ITU-T. P.800, "Methods for Subjective Determination of Transmission Quality," International Telecommunications Union, Geneva, Switzerland (1996).

[7] ITU-R Rep. BT.1082-1, "Studies toward the Unification of Picture Assessment Methodology," International Telecommunications Union, Geneva, Switzerland (1990).

[8] E. C. Poulton, *Bias in Quantifying Judgments* (Lawrence Erlbaum, London, 1989).

[9] A. Kohlrausch and S. van der Par, “Audio-Visual Interaction in the Context of Multi-Media Applications,” in *Communication Acoustics*, J. Blauert, Ed. (Springer, Berlin, Germany, 2005).

[10] D. Hands, “Multimodal Quality Perception: The Effects of Attending to Content on Subjective Quality Ratings,” in *Proc IEEE 3rd Workshop on Multimedia Signal Processing* (1999), pp. 503–508.

[11] S. Zieliński, F. Rumsey, S. Bech, B. de Bruyn, and R. Kassier, “Computer Games and Multichannel Audio Quality—The Effect of Division of Attention between Auditory and Visual Modalities,” presented at the AES 24th Int. Conf. (2003 May).

[12] M. H. Birnbaum, “Controversies in Psychological Measurements,” in *Social Attitudes and Psychophysical Measurement*, B. Wegener, Ed. (Lawrence Erlbaum, Hillsdale, NJ, 1982).

[13] H. T. Lawless and H. Heymann, *Sensory Evaluation of Food. Principles and Practices* (Kluwer-Plenum, London, 1998).

[14] M. Meilgaard, G. V. Civille, and B. T. Carr, *Sensory Evaluation Techniques* (CRC Press, New York, 1999).

[15] F. E. Toole, “Subjective Measurements of Loudspeaker Sound Quality and Listener Performance,” *J. Audio Eng. Soc.*, vol. 33, pp. 2–32 (1985 Jan./Feb.).

[16] S. Zieliński, “On Some Biases Encountered in Modern Listening Tests,” in *Proc. Int. Workshop on Spatial Audio and Sensory Evaluation Techniques* (Institute of Advanced Studies, University of Surrey, UK, 2006). Available from http://www.surrey.ac.uk/soundrec/ias/index.htm.

[17] G. A. Soulodre and M. C. Lavoie, “Subjective Evaluation of Large and Small Impairments in Audio Codecs,” presented at the AES 17th Int. Conf. “High-Quality Audio Coding” (1999 Aug.).

[18] U. B. Wüstenhagen, B. Feiten, T. Buholtz, R. Schwalve, and J. Kroll, “Method for Assessment of Audio Systems with Intermediate Audio Quality,” in *Proc. 21st Tonmeistertagung* (Hanover, Germany, 2003), pp. 655–671.

[19] ITU-R. BS. 1284, “Methods for the Subjective Assessment of Sound Quality—General Requirements,” International Telecommunications Union, Geneva, Switzerland (1997).

[20] F. E. Toole, “Listening Tests—Turning Opinion into Fact,” *J. Audio Eng. Soc.* (*Engineering Reports*), vol. 30, pp. 431–445 (1982 June).

[21] N. Zacharov and J. Huopaniemi, “Results of a Round Robin Subjective Evaluation of Virtual Home Theatre Sound Systems,” presented at the 107th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 47, pp. 1000, 1001 (1999 Nov.), preprint 5067.

[22] L. Gros, S. Chateau, and S. Busson, “Effects of Context on the Subjective Assessment of Time-Varying Speech Quality: Listening/Conversation, Laboratory/Real Environment,” *Acustica/Acta Acustica*, vol. 90, pp. 1037–1051 (2004).

[23] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson, “Recency Effect in the Subjective Assessment of Digitally-Codec Television Pictures,” in *Proc. 5th IEE Int. Conf. on Image Processing and Its Applications* (Edinburgh, UK, 1995), pp. 336–339.

[24] V. Seferidis, M. Ghanbari, and D. E. Pearson, “Forgiveness Effect in Subjective Assessment of Packet Video,” *Electron. Lett.*, vol. 28, pp. 2013–2014 (1992 Oct.).

[25] R. Hamberg and H. de Ridder, “Continuous Assessment of Perceptual Image Quality,” *J. Opt. Soc. Am.*, vol. 12, pp. 2573–2577 (1995).

[26] A. Watson, “Assessing the Quality of Audio and Video Components in Desktop Multimedia Conferencing,” Ph.D. thesis, Department of Computer Science, University College London (1999).

[27] S. P. Lipshitz and J. Vanderkooy, “The Great Debate: Subjective Evaluation,” *J. Audio Eng. Soc.*, vol. 29, pp. 482–491 (1981 July/Aug.).

[28] S. P. Lipshitz, “The Great Debate—Some Reflections Ten Years Later,” presented at the AES 8th Int. Conf., “The Sound of Audio” (1990).

[29] S. Bech, “Selection and Training of Subjects for Listening Tests on Sound-Reproducing Equipment,” *J. Audio Eng. Soc.*, vol. 40, pp. 590–610 (1992 July/Aug.).

[30] J. Blauert, *Communication Acoustics* (Springer, Berlin, Germany, 2005).

[31] S. Bech, “Training of Subjects for Auditory Experiments,” *Acta Acustica*, vol. 1, pp. 89–99 (1993).

[32] T. Neher, “Towards a Spatial Ear Trainer,” Ph.D. thesis, Institute of Sound Recording, University of Surrey, UK (2004).

[33] H. Fastl, “Psycho-Acoustics and Sound Quality,” in *Communication Acoustics*, J. Blauert, Ed. (Springer, Berlin, Germany, 2005).

[34] C. Guastavino, B. Katz, J. Polack, D. Levitin, and D. Dubois, “Ecological Validity of Soundscape Reproduction,” *Acustica/Acta Acustica*, vol. 91, pp. 333–341 (2005).

[35] T. Letowski, “Sound Quality Assessment: Cardinal Concepts,” presented at the 87th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 37, p. 1062 (1989 Dec.), preprint 2825.

[36] J. Blauert and U. Jekosch, “Sound Quality Evaluation—A Multi-Layered Problem,” *Acustica/Acta Acustica*, vol. 83, pp. 747–753 (1997).

[37] D. Västfjäll and M. Kleiner, “Emotion in Product Sound Design,” in *Proc. Journées Design Sonore* (Paris, France, 2002).

[38] F. E. Toole and S. Olive, “Hearing Is Believing vs. Believing is Hearing: Blind vs. Sighted Listening Tests, and Other Interesting Things,” presented at the 97th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 42, p. 1058 (1994 Dec.), preprint 3894.

[39] R. Bentler, D. Niebuhr, T. Johnson, and G. Flamme, “Impact of Digital Labeling on Outcome Measures,” *Ear and Hearing*, vol. 24, pp. 215–224 (2003).

[40] C. V. Beidl and W. Stücklschwaiger, “Application of the AVL-Annoyance Index for Engine Noise Quality,” *Acustica/Acta Acustica*, vol. 83, pp. 789–795 (1997).

[41] F. Rumsey, “Controlled Subjective Assessment of Two-to-Five-Channel Surround Sound Processing Algorithms,” *J. Audio Eng. Soc.*, vol. 47, pp. 563–582 (1999 July/Aug.).

[42] D. Västfjäll, “Contextual Influences on Sound Quality Evaluation,” *Acustica/Acta Acustica*, vol. 90, pp. 1029–1036 (2004).

[43] K. Zimmer, W. Ellermeier, and C. Schmid, “Using Probabilistic Choice Models to Investigate Auditory Unpleasantness,” *Acustica/Acta Acustica*, vol. 90, pp. 1019–1028 (2004).

[44] S. Choisel, “Spatial Aspects of Sound Quality,” Ph.D. thesis, Department of Acoustics, Aalborg University, Denmark (2005).

[45] R. Kirk, “Learning, A Major Factor Influencing Preferences for High-Fidelity Reproducing Systems,” *J. Audio Eng. Soc.*, vol. 5, pp. 238–241 (1957).

[46] K. Beresford, N. Ford, F. Rumsey, and S. Zieliński, “Contextual Effects on Sound Quality Judgments: Listening Room and Automotive Environments,” presented at the 120th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 54, p. 666 (2006 July/Aug.), convention paper 6648.

[47] K. Beresford, N. Ford, F. Rumsey, and S. Zieliński, “Contextual Effects on Sound Quality Judgements: Part II—Multistimulus vs. Single Stimulus Method,” presented at the 121st Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 54, pp. 1261, 1262 (2006 Dec.), convention paper 6913.

[48] B. A. Mellers and M. H. Birnbaum, “Loci of Contextual Effects in Judgment,” *J. Exp. Psychol.: Human Perception and Performance*, vol. 8, pp. 582–601 (1982).

[49] S. Zieliński, P. Hardisty, C. Hummersone, and F. Rumsey, “Potential Biases in MUSHRA Listening Tests,” presented at the 123rd Convention of the Audio Engineering Society, (*Abstracts*) www.aes.org/events/123/123rdWrapUp.pdf, convention paper 7179 (2007 Oct.).

[50] A. Parducci, “Contextual Effects: A Range-Frequency Analysis,” in *Handbook of Perception*, vol. 2, E. C. Carterette and M. P. Friedman, Eds. (Academic Press, London, 1974).

[51] A. Parducci, “Category Ratings: Still More Contextual Effects!” in *Social Attitudes and Psychophysical Measurement*, B. Wagner, Ed. (Lawrence Erlbaum, Hillsdale, NJ, 1982).

[52] H. Helson, *Adaptation-Level Theory* (Harper & Row, London, 1964).

[53] J. D. Harris and A. I. Rawnsley, “The Locus of Short Duration Auditory Fatigue or ‘Adaptation’,” *J. Exp. Psychol.*, vol. 46, pp. 457–461 (1953).

[54] F. E. Toole, “Loudspeakers and Rooms for Sound Reproduction—A Scientific Review,” *J. Audio Eng. Soc.*, vol. 54, pp. 451–476 (2006 June).

[55] J. P. Guilford, *Psychometric Methods* (McGraw-Hill, London, 1954).

[56] H. Stone and J. L. Sidel, *Sensory Evaluation Practices* (Academic Press, London, 1993).

[57] L. Bramsløw, “An Objective Estimate of the Perceived Quality of Reproduced Sound in Normal and Impaired Hearing,” *Acustica/Acta Acustica*, vol. 90, pp. 1007–1018 (2004).

[58] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach, “All Subjective Scales Are Not Created Equal: The Effects of Context on Different Scales,” *Signal Process.*, vol. 77, pp. 1–9 (1999).

[59] N. Narita, “Graphic Scaling and Validity of Japanese Descriptive Terms Used in Subjective-Evaluation Tests,” *SMPTE J.*, vol. 102, pp. 616–622 (1993 July).

[60] B. L. Jones and P. R. McManus, “Graphic Scaling of Qualitative Terms,” *SMPTE J.*, pp. 1166–1171 (1986).

[61] M. T. Virtanen, N. Gleiss, and M. Goldstein, “On the Use of Evaluative Category Scales in Telecommunications,” in *Proc. of Human Factors in Telecommunications* (Melbourne, Australia, 1995).

[62] K. Teunissen, “The Validity of CCIR Quality Indicators along a Graphical Scale,” *SMPTE J.*, pp. 144–149 (1996 Mar.).

[63] ITU-T P.910, “Subjective Video Quality Assessment Methods for Multimedia Applications,” International Telecommunications Union, Geneva, Switzerland (1999).

[64] R. Guski, “Psychological Methods for Evaluating Sound Quality and Assessing Acoustic Information,” *Acustica/Acta Acustica*, vol. 83, pp. 765–774 (1997).

[65] S. Zieliński, P. Brooks, and F. Rumsey, “On the Use of Graphic Scales in Modern Listening Tests,” presented at the 123rd Convention of the Audio Engineering Society, (*Abstracts*) www.aes.org/events/123/123rdWrapUp.pdf, convention paper 7176 (2007 Oct.).

[66] D. Clark, “High-Resolution Subjective Testing Using a Double-Blind Comparator,” *J. Audio Eng. Soc., Engineering Reports*, vol. 30, pp. 330–338 (1982 May).

[67] R. Conetta, “Scaling and Predicting Spatial Attributes of Reproduced Sound Using an Artificial Listener,” M.Phil./Ph.D. Upgrade Rep., Institute of Sound Recording, University of Surrey, UK (2007).

[68] S. K. Zieliński, F. Rumsey, and S. Bech, “Effects of Bandwidth Limitation on Audio Quality in Consumer Multichannel Audiovisual Delivery Systems,” *J. Audio Eng. Soc.*, vol. 51, pp. 475–501 (2003 June).

[69] D. C. Howell, *Statistical Methods for Psychology* (Wadsworth, UK, 2002).

[70] G. A. Miller, “The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information,” *Psych. Rev.*, vol. 63, pp. 81–97 (1956).

[71] R. Heusdens, J. Jensen, W. B. Kleijn, V. Kot, O. A. Niamut, S. Van De Par, N. H. Van Schijndel, and R. Vafin, “Bit-Rate Scalable Interframe Sinusoidal Audio Coding Based on Rate-Distortion Optimization,” *J. Audio Eng. Soc.*, vol. 54, pp. 167–188 (2006 Mar.).

[72] E. P. Köster, “The Psychology of Food Choice: Some Often Encountered Fallacies,” *Food Quality and Pref.*, vol. 14, pp. 359–373 (2003).

[73] L. B. Nielsen, “Subjective Assessment of Audio Codecs and Bitrates for Broadcast Purposes,” presented at

the 100th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 44, p. 634 (1996 July/Aug.), preprint 4175.

[74] S. S. Stevens, *Handbook of Experimental Psychology* (J. Wiley, London, 1951).

[75] N. H. Anderson, "Algebraic Models in Perception," in *Handbook of Perception*, vol. II (Academic Press, London, 1974).

[76] L. Narens, "A Theory of Ratio Magnitude Estimation," *J. Math. Psychol.*, vol. 40, pp. 109–129 (1996).

[77] K. F. Theusen, "Analysis of Ranked Preference Data," Master's thesis, Department of Informatics and Mathematical Modeling, Technical University of Denmark (2007).

[78] K. Zimmer and W. Ellermeier, "Deriving Ratio-Scale Measures of Sound Quality from Preference Judgments," *Noise Contr. Eng. J.*, vol. 51, pp. 210–215 (2003).

[79] S. Choisel and F. Wickelmaier, "Evaluation of Multichannel Reproduced Sound: Scaling Auditory Attributes Underlying Listener Preference," *J. Acoust. Soc. Am.*, vol. 121, pp. 388–400 (2007).

[80] ITU-R BT.500-10, "Methodology for the Subjective Assessment of the Quality of Television Pictures," International Telecommunications Union, Geneva, Switzerland (2000).

[81] R. Sperschneider, "Error Resilient Source Coding with Variable Length Codes and Its Application to MPEG Advanced Audio Coding," presented at the 109th Convention of the Audio Engineering Society, *J. Audio Eng. Soc.* (*Abstracts*), vol. 48, p. 1118 (2000 Nov.), preprint 5271.

THE AUTHORS

S. Zieliński

R. Rumsey

S. Bech

Slawek Zieliński holds a position of lecturer at the University of Surrey (UK). Previously he worked as a research fellow at the same university and as a lecturer at the Technical University of Gdansk (Poland). His current responsibilities include teaching electroacoustics, audio signal processing, and sound synthesis to undergraduate students. He is also responsible for the cosupervision of several Ph.D. students. He is in charge of the EPSRC-funded research project investigating new band-limitation strategies for 5.1 surround sound. He is also involved in a number of projects concerning topics such as quality evaluation of audio codecs, car audio optimization, and objective assessment of audio quality.

Dr. Zieliński is a member of the British Section of the Audio Engineering Society.

●

Francis Rumsey graduated in 1983 with first class honours (BMus Tonmeister) in music with applied physics, and subsequently worked with Sony Broadcast in training and product management. He was appointed a lecturer at Surrey in 1986 and received a Ph.D. degree from that university in 1991. He is professor and director of research at the Institute of Sound Recording, University of Surrey (UniS), and was a visiting professor at the School of Music in Piteå, Sweden, from 1998 to 2004. He was appointed a research advisor in psychoacoustics to NHK Science and Technical Research Laboratories, Tokyo, in 2006. He was a partner in EUREKA project 1653 (MEDUSA), studying the optimization of consumer multichannel surround sound. His current research includes a number of studies involving spatial audio psychoacoustics and he is currently leading a project funded by the Engineering and Physical Sciences Research Council, concerned with predicting the perceived quality of spatial audio systems, in collaboration with Bang & Olufsen and BBC Research.

Dr. Rumsey was the winner of the 1985 BKSTS Dennis Wratten Journal Award, the 1986 Royal Television Society Lecture Award, and the 1993 University Teaching and Learning Prize. He is the author of over 100 books, book chapters, papers, and articles on audio, and in 1995 he was made a fellow of the AES for his significant contributions to audio education. His book *Spatial Audio* was published in 2001 by Focal Press. He has served on the AES Board of Governors, was chair of the AES British Section, and was AES vice president, Northern Region, Europe. He was chair of the AES Technical Committee on Multichannel and Binaural Audio Technology until 2006.

●

Søren Bech received M.Sc. and Ph.D. degrees from the Department of Acoustic Technology of the Technical University of Denmark.

From 1982 to 1992 he was a research fellow there, studying the perception and evaluation of reproduced sound in small rooms. In 1992 he joined Bang & Olufsen as a technology specialist, and since 2007 he has been head of research. He is an adjunct professor at McGill University, Faculty of Music, and a visiting professor at the University of Surrey, Institute of Sound Recording. He has been project manager of several international collaborative research projects, including Archimedes (perception of reproduced sound in small rooms), ADTT (advanced digital television technologies), Adonis (image quality of television displays), LoDist (perception of distortion in loudspeaker units), Medusa (multichannel sound reproduction systems), and Vincent (flat panel display technologies).

Dr. Bech was cochair of ITU task group 10/3, member of ITU task group 10/4, and a member of the organizing committee and editor of a symposium on the perception of reproduced sound (1987). As a member of the AES Danish Section, he held numerous positions within the Audio Engineering Society, including AES governor and vice president, Northern Region, Europe. He is currently cochair of the AES Publications Policy Committee and on the review board of the *Journal* as well as the *Journal of the Acoustical Society of America* and *Acta Acustica*. He is a fellow of the AES and the Acoustical Society of America and he has published numerous scientific papers. He is the coauthor with N. Zacharov of *Perceptual Audio Evaluation—Theory, Method and Application* (Wiley, 2006).