Contextual Effects in Sensory Evaluation of Spatial Audio: Integral Factor or Nuisance?

William L. Martens[1]

[1] *Sound Recording Area, Schulich School of Music, McGill University, 555 Sherbrooke Street West, Montreal, QC, H3A 1E3, Canada*

Correspondence should be addressed to William L. Martens (wlm@music.mcgill.ca)

ABSTRACT

Experimental research has produced a great deal of data that is useful in predicting timbral and spatial attributes of the auditory imagery associated with reproduced sound. The model of sensory judgment upon which the majority of this research is based regards contextual effects as an unwanted nuisance that should be minimized in order to reveal underlying psychophysical relationships. A complementary view holds cognitive factors as integral to psychophysical measurement, and embraces the influence of context as a human factor worthy of study in its own right, and not simply in the interest of minimizing such influence. This paper will present these two complementary perspectives, and discuss their influences on experimental design within the field of preference testing for auditory spatial imagery associated with reproduced sound. A study designed to reveal potential differences in preference choices due to contextual effects is also presented to highlight the impact of design choices in which trial order is manipulated.

"In science we resemble children collecting a few pebbles at the beach of knowledge, while the wide ocean of the unknown unfolds itself in front of us."

- Sir Isaac Newton

1. INTRODUCTION

The development of techniques for sensory evaluation of spatial audio has been influenced greatly by psychological science, since such evaluation typically relies upon methods found in the discipline termed "psychological measurement." Because most advances in the application of these methods to sensory evaluation of spatial audio have been dependent upon the results of controlled experiments using human listeners, there is naturally a great concern over how best to design experiments that optimize the reliability and ultimate success of studies intended to examine human perception of spatial audio. A key factor to consider in this regard is the extent to which the sensory judgments that listeners make are dependent upon experimental context. If judgments are, then the results of the experiments in which listeners participate will exhibit context-influenced effects that call into question how generalizable the results will be.

Sir Isaac Newton's observation about the nature of science, which is quoted above in order to underscore this message, is a reflection on the immensity of the unknown in comparison to the tiny advances made by individual scientific studies. In the final analysis, the only knowledge that a given scientific study may provide without doubt is the knowledge of what data has been collected in that study. Generalizing beyond these data to what data might be collected in future studies, in other contexts, relies upon acceptance of assumptions and/or models that go beyond the data. And if an experimeter wishes to draw implications for practical applications of the results, then the contextual effects must be addressed most clearly. It should be clear that there is always a context for all studies, and that simply regarding

a contextual dependence as a nuisance to be eliminated does not begin to address the bigger problem posed here.

The motivation for this paper is to provide background on the study of contextual effects *per se*, and to provide instructive examples of how contextual effects may operate in perceptual studies of spatial sound reproduction, and how those contextual influences may be quantified in evaluations of spatial audio quality. It should be understood that one of the central goals in the sensory evaluation of spatial audio is simply to be able to predict whether one reproduced sound stimulus will be preferred to another. But if preferences depend upon context, it may be necessary to thoroughly understand contextual effects in sensory evaluation of spatial audio before listener responses can be successfully predicted on the basis of physical evaluations of reproduced sound fields. This paper attempts to clarify the influence of such contextual effects on reported preferences for auditory spatial imagery associated with reproduced sound, and present perspectives that hopefully will lead to a deeper understanding of the empirical phenomena of interest here.

Although recognizing the current need for a comprehensive model of how sensory features contribute to perceived quality of spatial sound reproduction, this paper is limited in scope by avoiding the discussion of this important endeavor (in much the same way as papers discussing the need for such a comprehensive model are often limited in scope by avoiding any substantial discussion of contextual effects). Nonetheless, this paper does discuss the concept of the "bottom-up" model of spatial image formation, in contrast to the "top-down" view of auditory spatial perception of complex scenes, as this paradigm shift in perceptual psychology is having a continuing influence upon how sensory evaluation techniques are currently evolving. In its traditional form, the "bottom-up" model of sensory judgment regards contextual effects as an unwanted nuisance that should be minimized in order to reveal underlying psychophysical relationships. Though this perspective might work in explaining basic psychophysical results, such as discrimination on clearly defined unidimensional attributes, there is often a need to incorporate consideration of the role of cognitive factors that allow for variation in a listener's responses each time they are presented with an identical stimulus.

A most elegant expression of these complementary perspectives appears in a recent book by Baird [1], who argues that some experimental outcomes are best attributed to sensory processes, while others are best attributed to judgment processes. Though the results of a minority of studies can be viewed equally well from either perspective, Baird questions whether any single model can explain the full spectrum of laboratory results, and argues that the ideal view would not attempt to support one model over another, but rather that these complementary perspectives must be taken together to provide the most adequate explanation of the phenomena under study:

> "In shorthand notation, this issue is often presented as a distinction between the influence of "sensory" and "cognitive" variables. A single model cannot accommodate both types of data, and it is time to frankly admit that attempts to explain all the facts from only one of these standpoints has failed."
>
> (Baird, [1], p. 2)

An example of a cognitive factor that is often a concern in preference tests for basic audio quality is that which allows stimulus order to influence the results of the test. Of course, this is a very common contextual effect even in psychophysical experiments on unidimensional attributes, such as loudness, and the operation of a so-called *sequential bias* within sensory judgment tasks has long been established [9]. The important point to underscore here is the following: When the order in which stimuli are presented makes a significant contribution to the results of an experiment, then response prediction based solely upon stimulus parameters measured within an isolated stimulus presentation will be relatively unsuccessful.

A concrete example can be found in historical studies of basic audio quality, where avoidable effects of stimulus order have hurt otherwise well executed listening tests: In violation today's standards, the

RM2 tests of MPEG-2/NBC [12] used the same sequence of items for all subjects, resulting in a clear contextual effect in which a stimulus presented after the "PitchPipe" stimulus got much worse scores than in subsequent tests that employed proper randomization of stimulus order.

In comparison to this *sequential bias* which is relatively easily fixed, however, there are more insidious contextual effects that can gradually shift for a given subject throughout the course of a single experimental session. The best known example of such is that which has been termed the *demand characteristic*. This term describes the experimental circumstances under which subjects may be induced to tell the experimenter what the experimenter wants to hear. While there is always a chance of an experimenter accidentally biasing listeners in some way that could be avoided, such is not the most common source of this undesireable influence on experimental results. Rather, it stems from a natural thought process that occurs in most listeners who contemplate what an experiment is about while they participate in it.

In the course of participating in an experiment, a listener might formulate an idea about what type of response the experimenter is expecting. This idea might come from interaction with the experimenter while the listener is receiving instructions, and so might not be included explicitly in the documented text of the instructions. Alternatively, the idea might develop during the execution of a sensory judgment task, being sensitive to the experimental situation, or from prior experience that leads the listener to imagine what the experiment might be examining. Whether the listener gets the right idea about the experiment or not, it is the formulation of the idea of what is expected that can more or less strongly influence what responses the listener makes, which may be different from how they might respond if they were listening in some other context.

It should be clear that this is not just a problem with research methods, but a problem with the model of how human listeners form auditory spatial imagery. That is, listeners' expectations play a significant role not only in how they will report on their auditory spatial perceptions, but also in what perceptions will likely be formed given the perceptual hypotheses they anticipate testing. Ever since Neisser's [6] perceptual cycle was proposed as an explanation of how knowledge, perception, action, and the environment all interact in goal-oriented human behavior, it has become more common to question the "bottom-up" model of spatial image formation. In most listening experiments related to spatial quality, listeners are likely to be given the opportunity to explore their perceptions of the presented auditory scene, either before or during experimental trials (see Rumsey [10] for a proposal of a scene-based paradigm for the evaluation of spatial sound reproduction).

Neisser's perceptual cycle assumes a reciprocal relationship between the listener's schema (i.e., knowledge about the environment) and actions (i.e., active explorations of the auditory scene). The listener's active listening determines what information will be picked up from the environment, which information in turn modifies the listener's schema. A clear statement about this paradigm shift towards a more "top-down" view of auditory spatial perception is found in the recently published edition of Blauert's book on "Spatial Hearing." In the newly added section entitled "Progress and Trends since 1982" he wrote about the classic "bottom-up" view:

> "It was understood that the signal processing...is basically 'signal-driven', that is, the input signals to the two ears essentially determine the resulting representation of binaural activity... Yet the ultimate and relevant output of the auditory systems is not any hypothetical internal representation of binaural activity, but rather auditory perceptual scenes."
>
> (Blauert, [2], p. 409)

Then, regarding the "top-down" view he wrote:

> "Pattern recognition is a 'hypothesis-driven' process. At a given moment in time, the system typically sets up the hypothesis that a certain pattern of attributes is contained in the data. This hypothesis is then checked and subsequently accepted or rejected. In terms of information flow, such a system shows a so-called 'top-down' architecture."
>
> (Blauert, [2], p. 410)

So, if it is accepted that listeners' expectations play a significant role in both the formation auditory spatial images, and in how listeners will report on those images, then it is critically important to attempt to control the experimental context in all of its components, including setting, instructions, feedback, etc. As will be explained in greater depth below, the problem of how trials are blocked for comparison is a particularly influential detail. This issue would not come up at all if listening experiments related to spatial audio quality could be executed so that only a single judgment were collected from each listener, and then a new listener were randomly chosen to judge the next stimulus. Besides the impracticality of such a *between-subject* design, it is also not nearly as powerful as a listening test that employs repeated measures. In fact, it is often possible for an entire experiment to be completed according to a *within-subject* design, where each subject provides data under many different conditions, receiving all treatments that other subjtects receive, and giving responses for all stimuli presented in the experiment. If this is done, then stimulus ordering can be arranged so that contexually-based *nuisance variables* are handled in a potentially informative way, rather than handling them in a way intended simply to minimize their influence.

When a *nuisance variable* is included in a study explictly, as an integral factor, it is usually distinguished from other random variables through use of the term *nuisance factor*. It is one of the goals of this paper to elevate the *nuisance factor* to a higher status in the community of researchers who are active in experimental design of listening tests. This special status is in contrast to classic definition of the term, which has been formulated as follows: A *nuisance factor* is "a variable in which the experimenter has no real interest but cannot actually be ignored." [5]. A more powerful perspective on how to deal with the *nuisance variable* / *nuisance factor* distinction is found in the excellent researcher's handbook on design and analysis by Keppel & Wickens [3], who teach about balancing carefully the strategies employed to handle *nuisance variables*. There are four common ways to deal with them, and many experiments are designed using combination of these four:

- Hold a *nuisance variable* to a constant value throughout an experiment;
- *Counterbalance* a *nuisance variable* by including all of its values equally often in each condition of an experiment;
- Include a *nuisance variable* as an explict factor in an experiment;
- Destroy the systematic relationship between a *nuisance variable* and an *independent variable* through randomization.

Especially because of problems with the *confounding* of the effects that a *nuisance variable* and an *independent variable* have on a *dependent variable*, randomization is the most common solution. However, the potential importance of contextual effects in understanding the application of experimental results argues for a more sensitive approach that might avoid *confounding* while treating *nuisance variables*:

> "In any study, there is an infinitude of potential nuisance variables, some of them important, most not. In many studies all four strategies [above mentioned] are employed. An index of the skill of a researcher is the subtlety with which this is done."
>
> (Keppel & Wickens, [3], p. 6)

Because manipulation of experimental design plays a role in the discussion of the experiment that follows in the next section, it would be best to reiterate the points just made: In *between-subject* designs, the systematic relationship between a *nuisance variable* and an *independent variable* is often handled by random assignment of subjects to conditions in order to minimize chances for *confounding*. But since listening tests typically use repeated measures, this simple solution is not available, and so the context within which each stimulus is presented for a given subject becomes important. A *nuisance variable* may be included explicitly in the design of an experiment so that its interaction with an *independent variable* can be observed in all combinations. When possible, this will be done by *counterbalancing* the *nuisance variable* so that each of its levels occur equally often at all levels of an *independent variable*, hopefully nullifying its effect on observed mean values of a *dependent variable* in their dependence upon that *independent variable*.

2. A STUDY OF CONTEXTUAL EFFECTS

The study of contextual effects to be described here had two primary goals. One goal was to determine whether different multichannel microphone techniques might be preferred for recording and reproduction of different types of musical performance. A second goal was to determine the experimental context within which such preferences might be observed, which it was hypothesized might depend upon stimulus ordering. For this reason, listeners were split into two groups who would be given different stimulus ordering, but would nonetheless hear the same stimuli and make the same comparisons between them. The details of the study are described elsewhere [4]. What is to be presented here focusses on the experimental design for the study, and how the results were found to indicate a strong contextual effect on preference choices regarding spatial audio quality. Only a brief introduction to the study is provided here.

2.1. Reproducing the Spatial Imagery of Piano Performances

Four solo piano pieces composed in the European concert music tradition and deemed to be representative of different eras were recorded using four different surround microphone arrays.

- Bach
- Schumann
- Brahms
- Contemporary improvisation (Tom Plaunt)

Each microphone array was positioned in order to optimize its perceived sound quality. The resulting multichannel sound reproductions were approved by several professionals with *Tonmeister training*. Although all sixteen recordings sounded quite good, they differed in terms of the spatial imagery they presented to the listener (and steps were taken to ensure that reproductions were subjectively well matched in timbre). Though no discussion of microphone techniques will be included in this paper, for the sake of the interested reader, the following list of the employed surround microphone arrays is provided:

- Fukada tree
- Polyhymnia Pentagon (5 omnis)
- OCT + Hamasaki Square
- SoundField

2.2. Preferences for the Piano Performance Reproductions

After an informal listening session in which participants were allowed to hear all four versions of the four piano performances, each listener completed four blocks of paired comparisons in which the task for each trial was to choose which of the two versions of a single performance was the prefered version. For example, on a given trial, a listener might be presented with an excerpt of the Bach piece recorded either with the Fukada tree or the Polyhymnia Pentagon. No indication was ever given regarding which microphone technique was associated with which stimulus presentation, as the two stimuli were labelled *A* and *B*. After listening for as long as they desired, with free switching between versions allowed, listeners indicated their preference and moved on to the next trial. Identical versions were never presented for comparison, and each comparison between two versions for a given piece was made twice, with a reversal of microphone techniques labeled *A* and *B*. All listeners completed four blocks of 12 preference choices, so that all heard all combinations of microphone techniques and musical selections; however, the order in which the trials were completed differed between listeners.

2.3. Successive versus Intermixed Trial Ordering

Two groups were formed, each containing 18 musically experienced listeners, and these two groups completed trials according to two different trial ordering schemes. The pairwise-comparison trials themselves were identical, as were the instructions that the subjects were given; however, for one group all trials for a given musical selection were completed in a single block, and then the experiment progressed to a block of trials for a different musical selection. This approach to trial ordering has been termed the *successive-treatment design* [3]. The second group of 18 listeners also completed four blocks of 12 preference-choice trials, but the musical selection

was randomly assigned from trial to trial, so that the presentation of the four musical selections was distributed throughout the 48 trials. This approach to trial ordering has been termed the *intermixed-treatment design* [3]. Thus, for this group of 18 listeners, any effects due to sequential biases might be likely to be nullified, since the trial order was different for each listener. In contrast, the group of listeners receiving successive trial ordering had trial order randomized only within blocks of 12 trials, rather than over the entire 48 trials. Of course, in such a *successive-treatment design*, the order of the blocks of single-musical-selection trials is a matter for concern. Therefore, the order in which the successive four blocks were completed was also randomized for the listeners in the successive-trial-ordering group.

2.4. Influence of Musical Selection on Preference

The two groups of listeners who completed 48 otherwise identical pairwise-comparison trials, gave a different pattern of preference choice results. The successive-trial-ordering group preferred both the Fukada tree and the Polyhymnia Pentagon to nearly the same extent. That is, either of these two could be used as a good microphone technique for recording all four musical selections, regardless of whether the performance was of a composition by Bach, Schumann, or Brahms, or a contemporary improvisation. In contrast, for the intermixed-trial-ordering group the pattern of preference choices was strongly influenced by musical selection (as indicated by statistical tests reported in [4]). How can these contrasting results be explained? Without examining the detailed results further, it is relatively easy to propose a satisfying explanation.

When listeners are presented with the same musical selection again and again within a single block of trials, they quickly come to focus upon the particular differences between versions of that musical selection that result from the use of different microphone techniques. These are differences in the spatial imagery *for that musical selection*. That this could happen within the duration of a block of 12 trials is perhaps a bit surprising, but all listeners in this study were experienced listeners.

When listeners are presented with the different musical selections on each trial, they focus not upon the particular differences between versions for a given musical selection, but rather maintain a more global perspective on all the spatial imagery they hear across trials. This would allow them to express their preference for the Fukada tree as producing the best spatial image for one musical selection, while preferring the Polyhymnia Pentagon microphone technique for recording a different piano performance.

2.5. Generalization

If an experiment is intended to test for the influence of musical selection on preferred microphone technique, it is critically important to structure the trials in order to allow such a preference to be exhibited. On the other hand, if an answer is desired for the more general question about microphone technique, when an experimenter wishes to improve chances for generalizing beyond the particular set of musical selections employed, then, perhaps counterintuitively, it might be best to highlight distinctions made by microphone techniques for a particular musical performance. This somewhat counterintuitive finding of a stimulus ordering effect is not without precedent. Olive, et al. [7] obtained a similar result in a study of the influence of room acoustics on preferences for loudspeakers, though perhaps for other reasons. It is instructive to examine that study in some of its details to find where it overlaps with the current study, and where it may differ.

First it should be pointed out that the study by Olive, et al. [7] collected preference ratings rather than preference choices, and used a multistimulus comparison rather than the pairwise comparison trial structure employed here. But an analogous dependence upon context was observed in terms of how preferences were expressed. Although loudspeaker position was also a variable in that study, the primary stimulus components to be compared were speakers (of which there were three) and rooms (of which there were four). To allow for rapid comparison between reproduced sound in the four different rooms, all stimuli were recorded binaurally. The most straightforward story here can be found in the comparison of preference ratings between two types of multistimulus-comparison trials. One could be termed "within-room," since three speakers were compared within a single room in each trial, with rooms varying over trials. The other could be termed "among-rooms," since the acoustic influence of four

reproduction rooms on a single loudspeaker in a single position was observed in each trial, with loudspeaker varying over trials. What happened was that the loudspeaker had a statistically significant effect upon preferences in the "within-room" condition, and room did not. The opposite results were obtained in the "among-rooms" condition, with room having the largest effect on preferences. What they concluded from these results is the following:

> "These contrasts in loudspeaker and room effects indicate that subjective measurements of sound quality are relative measurements strongly biased by the context in which the measured objects are compared."
>
> (Olive, et al. [7], p. 1)

There is an additional methodological aspect of the study by Olive, et al. [7] that is worth contemplating here, since the multistimulus-comparison trials had a different composition in the "within-room" and the "among-rooms" conditions. It might be that when listeners are presented with comparisons between loudspeakers within a room, they acclimate to the room acoustic environment, and learn to ignore these influences while they focus only upon loudspeaker characteristics. In order to support the conclusion that this acclimation is occurring, it might be better to hold the trial composition constant for two groups of listeners, as was done for the piano performances in the current study. Olive [8] has indicated that in a future related study, each multistimulus-comparison trial might compare different loudspeakers within a single room on each trial, but the room might be held constant within a block of trials for one group of listeners, and completely intermixed across trials for the other group.

Just in closing this section, it is worth commenting upon where the strongest contextual dependencies are likely to be observed in such preference tests. The difference between the two tasks discussed here, the pairwise-preference-choice task versus the multistimulus-rating task, might lead one to suppose that the listener must be required to cognize to a greater extent when comparing multiple stimuli on a continuous preference scale, rather than when simply making a choice between two stimuli. It would seem that this would lead to a greater chance for some types of contextual dependencies to exert their influence on the resulting listener behavior. Suffice it to say, however, that contextual effects come in all shapes and sizes, as is underscored in the following conclusion, and its associated quotation.

3. CONCLUSION

This paper attempted to clarify the influence of contextual effects on sensory evaluation of auditory spatial imagery associated with reproduced sound. The study of contextual effects reported here addressed one of the central goals in the sensory evaluation of spatial audio, which is simply to be able to predict whether one reproduced sound stimulus will be preferred to another. A simple statement that preferences will always depend upon context does not do justice to the subtlety and complexity of how preferences depend upon context, which is a topic worthy of deep investigation.

With regard to spatial audio quality, it must be pointed out that it can be no more absolute than is timbral quality. To say that a listener's sensory judgments about these multidimensional attributes are relative, is another way of saying that they depend upon context. When the order in which stimuli are presented makes a significant contribution to the results of an experiment, then response prediction based solely upon stimulus parameters measured within an isolated stimulus presentation will be relatively unsuccessful. Such cases reveal the complexity of human judgment processes, and argue for careful consideration of the role of learning and memory in experimental task construction. Indeed, the way in which listeners may adapt to stimulus context throughout the course of an experiment argues for the design of experiments that target the quantification of biases based upon stimulus order and stimulus range, and that examine the role that instructions play in determining the strategies listeners take in performing experimental tasks.

There is always a temptation to end a paper such as this with a provocative statement. Just as what is heard by an individual may depend on those sensory attributes for which they are listening, so may what results are found by a researcher depend upon their presuppositions regarding what might be

found. This will depend upon personal perspectives taken on what the highest priority experimental questions are in a particular context. Since it may be that the design choices made in constructing new listening experiments will always be based upon what an experimenter presumes are the best hypotheses to test, one could suppose that details, such as the stimulus and trial ordering, will be influenced by the experimenter's biases. When these biases operate unconsciously, they can lead to poor choices, with the potential for developing misleading results, like the finding that musical selection makes no difference to preferences for microphone techniques. When such biases are themselves of interest, however, there may be a way for them to be included in the design of a study as testable hypotheses. Either way, when formulating scientific experiments for the evaluation of spatial audio, the following warning is worth contemplating:

> "... the presuppositions of science are normally mistaken for its findings."
>
> - E. F. Schumacher, [11], p. 94

4. ACKNOWLEDGMENT

This research was supported in part by *Valorisation-Recherche Québec* (VRQ) of the Government of Quebec within the project *Real-time Communication of High-resolution Multi-sensory Content via Broadband Networks*, and by the Centre for Interdisciplinary Research in Music Media Technology (CIRMMT).

5. REFERENCES

[1] J. C. Baird. *Sensation and Judgment: Complementarity Theory of Psychophysics.* Lawrence Erlbaum Associates, Mahwah, N.J., 1997.

[2] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization, (Revised Edition).* MIT Press, Cambridge, Massachusetts, 1997.

[3] Geoffrey Keppel and Thomas D. Wickens. *Design and Analysis: A Researcher's Handbook.* Pearson Education, Upper Saddle River, New Jersey, fourth edition, 2004.

[4] Sungyoung Kim, Martha DeFrancisco, Kent Walker, Atsushi Marui, and William L. Martens. Listener preferences in multichannel audio: Examining the influence of musical selection on surround microphone technique. In *Proceedings of 28th International Conference (to appear)*, Piteå, Sweden, July 2006. Audio Eng. Soc.

[5] J. Kurtz. Glossary. Web site, September 1997. `http://www.itl.nist.gov/iaui/vvrg/nist-icv/experiments/mapnav/dev/node4.html`.

[6] U. Neisser. *Cognition and Reality: Principles and Implications of Cognitive Psychology.* W H Freeman & Co, New York, N. Y., 1976.

[7] S. E. Olive, P. L. Schuck, S. L. Sally, M. E. Bonneville, K. L. Momtahan, and E. S. Verreault. The variability of loudspeaker sound quality among four domestic-sized rooms. In *Proceedings of 99th Convention.* Audio Eng. Soc., October 1995. Preprint 4092.

[8] Sean Olive. personal communication, 2005.

[9] E. C. Poulton. *Bias in Quantifying Judgments.* Lawrence Erlbaum Associates, London, England, 1989.

[10] Francis Rumsey. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of Audio Engineering Society*, 50(9):651–666, September 2002.

[11] E. F. Schumacher. *Small Is Beautiful : Economics as if People Mattered.* Hartley & Marks Publishers, Vancouver, B.C., 1999.

[12] Thomas Sporer. personal communication, 2005.