OPEN ACCESS Freely available online
Journal of the Audio Engineering Society
Vol. 64, No. 6, June 2016 (© 2016)
DOI: http://dx.doi.org/10.17743/jaes.2016.0015

A Meta-Analysis of High Resolution Audio Perceptual Evaluation

JOSHUA D. REISS, *AES Member*
(joshua.reiss@qmul.ac.uk)

Queen Mary University of London, London, UK

There is considerable debate over the benefits of recording and rendering high resolution audio, i.e., systems and formats that are capable of rendering beyond CD quality audio. We undertook a systematic review and meta-analysis to assess the ability of test subjects to perceive a difference between high resolution and standard, 16 bit, 44.1 or 48 kHz audio. All 18 published experiments for which sufficient data could be obtained were included, providing a meta-analysis involving over 400 participants in over 12,500 trials. Results showed a small but statistically significant ability of test subjects to discriminate high resolution content, and this effect increased dramatically when test subjects received extensive training. This result was verified by a sensitivity analysis exploring different choices for the chosen studies and different analysis approaches. Potential biases in studies, effect of test methodology, experimental design, and choice of stimuli were also investigated. The overall conclusion is that the perceived fidelity of an audio recording and playback chain can be affected by operating beyond conventional levels.

1 INTRODUCTION

High resolution audio may be loosely defined as those systems and formats that are capable of rendering beyond standard quality audio, i.e., more than 16 bits, and/or more than 44.1 or 48 kHz sample rate, as used in Compact Disc (CD) or "ordinary" Digital Video Disc (DVD) quality audio. Yet many believe that this standard quality audio is sufficient to capture all perceivable content from live sound. This question of perception of high resolution audio has generated heated debate for many years. Although there have been many studies and formal arguments presented in relation to this, there has yet to be a rigorous analysis of the literature.

By analyzing the data from multiple studies, it should be possible to come up with more definitive results concerning the perception of high resolution audio. For instance, several tests used similar methodologies and so it might be possible to pool the data together. In other cases, data is provided on a per subject level, which could allow re-analysis.

Here, we provide a meta-analysis of those studies. Note that this is far more than a literature review, since it compiles data from multiple studies, performs statistical analyses on this aggregate data, and draws new conclusions from the results of this analysis. Meta-analysis is a popular technique in medical research and has been applied to the evaluation of music information retrieval techniques [1–3]. The term has also been applied to primary analysis of the performance of audio feature extraction techniques within a general framework [4]. But to the best of our knowledge, this represents the first time that it has been applied to audio engineering research.

1.1 Reviews

There are several overviews of the field of high resolution audio relevant to this work. A special issue of the *Journal of the Audio Engineering Society* was dedicated to the subject [5], although none of the papers therein was focused on the question of perception. [6–9] all gave detailed descriptions of suggested requirements for high resolution audio formats and systems. [10, 11] provided reviews of high resolution audio perceptual evaluation. [12] gives guidelines and recommendations for high resolution audio listening tests. Together, these works serve as an excellent introduction to the subject and the important research questions.

[13] provided a systematic review of studies concerning the health effects of exposure to ultrasound. The studies reviewed showed that it may be associated with hearing loss, dizziness, loss of productivity, and other harmful effects. However, some of the reviewed studies defined ultrasound as beyond 10 kHz, thus including content known to be audible. And all studies discussed in [13] focused on prolonged exposure, especially in the work environment.

1.2 Identification and Selection of High Resolution Audio Studies

In total, 80 relevant references pertaining to high resolution audio perception were found from which we identified 18 experiments suitable for meta-analysis. This section describes the search methods used to identify relevant research, as well as the selection criteria for inclusion or exclusion of studies in the secondary and meta-analysis.

The review papers mentioned in the previous section may be considered the starting point for this work. We searched through all references they cited and all papers that have cited any of them in order to identify any relevant experiments. For all of the papers identified that concerned perception of high resolution audio, we then repeated the procedure, searching all citations therein and all citations of those papers. This procedure was repeated until no new potentially relevant references could be found. Potentially relevant experiments were also found based on discussions with experts, keyword searches in databases, and search engines and the author's prior knowledge. The same iterative search on the citations within and citations of those papers was again applied to these additional papers. In total, 80 relevant references were found, of which there were 51 papers describing perceptual studies of high resolution audio.

No experiments published before 1980 were considered. A study of potentially relevant references showed that they mainly assumed that content beyond 20 kHz would be unnecessary and may not have had sufficiently high quality equipment to reproduce high resolution audio anyway [14–21].

Several potentially relevant references could not be found. These were all non-English language publications. Furthermore, they were often presentations in meetings and so may not have been formally published. But in all cases, the authors had English language publications and it appeared that the English language versions may have described the same experiment.

There may also be relevant experiments that were overlooked because they had an unusual methodology, were described in an unusual way or presented to a very different audience. This is most likely the case for works published in physics or neuroscience journals. However, for all the relevant experiments that were found described in such places, though they dealt with aspects of high resolution audio, they did not focus directly on the most fundamental questions with which we are concerned, that is, the discrimination between standard quality and beyond standard quality audio with real world content.

Many publications treated results for different conditions, such as different stimuli or different filters for sample rate conversion, as different experiments. Since these experiments generally have the same participants, same investigators, same methodology, etc., they were grouped as a single study. Where the experiments involved fundamentally different tasks, as in [22–24], these were treated as different studies.

Studies focused on auditory perception resolution were not considered. Such studies may suggest the underlying causes of high resolution audio discrimination, if any, but they are not directly focused on discrimination tasks. Similarly, experiments involving indirect discrimination of high resolution audio were excluded because an indirect effect may be observed or not, regardless of whether high resolution audio can be directly discriminated. In particular, brain response to high resolution content may not even relate to perception.

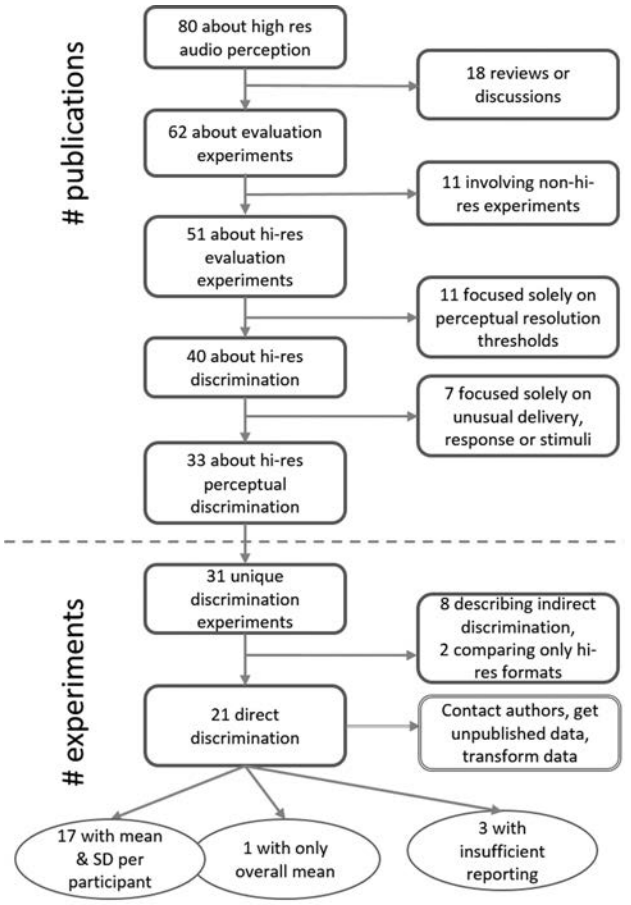

Fig. 1. Flowchart highlighting the study selection process.

Studies focused on discrimination between competing high resolution formats, or on discrimination when only low resolution content is used, are not applicable since they either don't address detecting a difference between those formats and standard resolution, or intentionally don't use content that would effectively demonstrate such a difference.

Within the studies focused on perceptual discrimination, we identified at least 21 distinct, direct discrimination studies. Three of these [25–27] were excluded because there was insufficient or unusual reporting that would not allow use in meta-analysis. Fig. 1 presents a study flow diagram showing how the studies were selected for meta-analysis.

1.3 Classification of High Resolution Audio Studies

Table 1 provides a near complete listing of all perceptual studies (i.e., listening tests) involving high resolution audio. Studies generally are divided into those focused on

Table 1. Summary and classification of high resolution audio listening tests.

Test Type			Reference	Methodology
Auditory perception resolution	Bone conduction perception		[28–30]	pattern recognition, frequency JND
	Temporal resolution		[31]	2IFC
			[32–34]	ABX
	Frequency resolution		[35]	Method of limits
			[34, 36–40]	2IFC
	Joint time-frequency resolution		[41, 42]	2IFC
Format discrimination	Indirect discrimination	Brain response	[22, 39, 40, 43–51]	N/A
		Semantic description	[22, 43–46, 48, 52, 53]	DoD, Attribute rating
		Other (level, spatialisation, temporal resolution)	[22, 23, 45, 46, 51, 54, 55]	Method of adjustment; Method of limits
	Sufficient formats discrimination	Alternative Hi-Res Formats	[56]	ABX
			[57]	AB
		Low resolution content	[10]	AXY
			[58]	Same different
	High vs standard discrimination	Test signals	[59]	Same different
			[25, 26, 60, 61]	2IFC
		Real world content	[11, 24, 35, 43, 58, 62–65]	Same different
			[23, 66]	ABX
			[64, 67–69]	AXY
			[24, 70]	XY
			[71, 72]	Multistimulus rating

establishing the limits of auditory perception and those focused on our ability to discriminate differences in format.

1.3.1 Auditory Perception Resolution Studies

In the former category, several studies have focused on bone conduction, where the transducer is placed in direct contact with the head, e.g.,[28]. This assisted form of rendering high resolution audio does not correspond with typical listening conditions, though it is possible that bone conduction may assist perception over headphones.

However, the majority of perceptual resolution studies have been concerned with time and frequency resolution. A major concern is the extent to which we hear frequencies above 20 kHz. Though many argue that this would not be the primary cause of high resolution content perception, it is nevertheless an important question. [36, 37, 39, 40] have investigated this extensively, and with positive results, although it could be subject to further statistical analysis.

Temporal fine structure [73] plays an important role in a variety of auditory processes, and temporal resolution studies have suggested that listeners can discriminate monaural timing differences as low as 5 microseconds [31–33]. Such fine temporal resolution also indicates that low pass or anti-alias filtering may cause significant and perceived degradation of audio when digitized or downsampled [54], often referred to as *time smearing* [74]. This time smear, which occurs because of convolution of the data with the filter impulse response, has been described variously in terms of the total length of the filter's impulse response including pre-ring and post-ring, comparative percentage of energy in the sidelobes relative to the main lobe, the degree of pre-ring only, and the sharpness of the main lobe.

[41, 42] both claim that human perception can outperform the uncertainty relation for time and frequency resolution. This was disputed in [75], which showed that the conclusions drawn from the experiments were far too strong.

1.3.2 Format Discrimination Studies

Studies in this category are in some sense focused on our ability to discriminate the rendering of high resolution content or formats. Many of these studies may be considered indirect discrimination, since they don't ask participants to select a stimuli or to identify whether a difference exists. Notable among these are studies that measure brain response. [44] showed that high frequency sounds are processed by the brain and observed an increase in activity when listeners were presented with broad-spectrum signals compared with those containing either the low frequency signal (below 22 kHz) or high frequency signal (above 22 kHz) alone. But this does not necessarily imply that high resolution audio is consciously, or even subconsciously, distinguished.

Other forms of indirect discrimination include studies that ask participants to identify or rate semantic descriptors [44, 52], or to perform a task with or without high resolution audio, e.g., localize a sound source [23], set listening level [46], discriminate timing [54]. Such studies may show, at a high level, what perceptual attributes are most affected. However, the difficulty with subjecting such

studies to meta-analysis is that a well-designed experiment may (correctly) give a null result on the indirect discrimination task even if participants can discriminate high resolution audio by other means.

Several studies have been focused on tasks involving direct discrimination between competing high resolution audio formats. In [56], test subjects generally did not perceive a difference between DSD (64×44.1 kHz, 1 bit) and DVD-A (176.4 kHz, 16 bit) in an ABX test, whereas [57] showed a statistically significant discrimination between PCM (192 kHz/24 bits) and DSD. However, in both cases, high resolution audio formats are compared against each other. Certainly in the first case, the null result does not suggest that there would be a null result when discriminating between CD quality and a higher resolution format. The second case is intriguing, but closer inspection of the experimental set-up revealed that the two formats were subject to different processing, most notably, different filtering of the low frequency content.

2 SECONDARY ANALYSIS

Table 2 A lists the studies that were included in the secondary analysis and meta-analysis. For the remainder of the paper, they are referred to by "AuthorYear" notation, to distinguish the studies from related publications (many studies were described in multiple publications, and some papers described multiple studies). In this section we revisit data from these studies, where available, in order to perform additional analysis of the results and to present the results in a form suitable for later meta-analysis.

2.1 Transformation of Study Data

Yoshikawa 1995 involved discrimination of 96 kHz and 48 kHz in an AXY test. Although only t values are reported for each stimulus/participant combination, these are derived from trials with a discrete set of results. By computing all possible sets of results and comparing the resultant t values with the reported t values, we were able to estimate the number of correct answers for each participant.

In King 2012, participants were asked to rate 44.1 kHz, 96 kHz, 192 kHz, all at 24 bit, and "live" stimuli in terms of audio quality. This methodology is problematic in that the ranking may be inconclusive, yet people might still hear a difference, i.e., some may judge low sample rate as higher quality due to a personal preference, regardless of their ability to discriminate.

We were provided with the full data from the experiment. A priori, the decision was made to treat the "live" stimuli as a reference, allowing the ranking data to be transformed into a form of A/B/X experiment. For each trial, it was treated as a correct discrimination if the highest sample rate, 192 kHz, was ranked closer to "live" than the lowest sample rate, 44.1 kHz, and an incorrect discrimination if 44.1 kHz was ranked closer to "live" than 192 kHz. Other rankings were excluded from analysis since they may have multiple interpretations. Thus if there is an inability to discriminate high resolution content, the probability of a correct answer is 50%.

In Repp 2006, participants also provided quality ratings, in this case between 24 bit / 192 kHz, 16 bit / 44.1 kHz, and lower quality formats. This can be transformed into an XY test by assuming that correct discrimination is made when 24 bit / 192 kHz was rated higher than 16 bit / 44.1 kHz, and incorrect discrimination if 24 bit / 192 kHz was rated lower than 16 bit / 44.1 kHz. Results where they are rated equal are ignored, since there is no way of knowing if participants perceived a difference but simply considered it too small compared to differences between other formats, and hence cannot be categorized. Note also that here, unlike King 2012, there is no reference with which to compare the high resolution and CD formats. Thus, without training, there may be no consistent definition of quality and it may not be possible to identify correct discrimination of formats.

2.2 Meyer 2007 Revisited

Meyer 2007 deserves special attention, since it is well-known and has the most participants of any study, but could only be included in some of the meta-analysis in Sec. 3 due to lack of data availability. This study reported that listeners could not detect a difference between an SACD or DVD-A recording and that same recording when converted to CD quality. However, their results have been disputed, both in online forums (www.avsforum.com, www.sa-cd.net, www.hydrogenaud.io and secure.aes.org/forum/pubs/journal/) and in research publications [11, 76].

First, much of the high-resolution stimuli may not have actually contained high-resolution content for three reasons; the encoding scheme on SACD obscures frequency components above 20 kHz and the SACD players typically filter above 30 or 50 kHz, the mastering on both the DVD-A and SACD content may have applied additional low pass filters, and the source material may not all have been originally recorded in high resolution. Second, their experimental set-up was not well-described, so it is possible that high resolution content was not presented to the listener even when it was available. However, their experiment was intended to be close to a typical listening experience on a home entertainment system, and one could argue that these same issues may be present in such conditions. Third, their experiment was not controlled. Test subjects performed variable numbers of trials, with varying equipment, and usually (but not always) without training. Trials were not randomized, in the sense that A was always the DVD-A/SACD and B was always CD. And A was on the left and B on the right, which introduces an additional issue that if the content was panned slightly off-center, it might bias the choice of A and B.

Meyer and Moran responded to such issues by stating [76], "... there are issues with their statistical independence, as well as other problems with the data. We did not set out to do a rigorous statistical study, nor did we claim to have done so. ..." But all of these conditions

Table 2. A. List of studies included in meta-analysis. B. Risks of potential biases and other issues in the included studies (see Sec. 2.5). Low risk "-"; unclear risk "?"; high risk "□." The last column identifies if these risks tend strongly towards false positives (Type I errors), false negatives (Type II errors) or neither (Neutral). C. Total number of trials and correct answers for each study, with the associated binomial probability (see Sec. 3.1).

A. Study			B. Risk of Bias							C. Binomial test			
Study	Year	Main references	Allocation bias	Methodology	Experimental design	Stimuli bias	Attrition bias	Reporting bias	Leading to	# correct	total	percent correct	probability
Plenge	1980	[59]	–	?	□	□	–	–	Type II errors	1367	2580	52.98%	**1.294E-03**
Muraoka	1981	[35]	–	?	□	?	–	–	Neutral	542	1060	51.13%	0.2400
Oohashi	1991	[43]	–	–	–	–	–	?	Neutral	392	800	49.00%	0.7261
Yoshikawa	1995	[67]	–	–	?	?	?	–	Neutral	85	132	64.39%	**5.976E-04**
Theiss	1997	[23]	–	?	?	?	□	–	Neutral	38	51	74.51%	**3.105E-04**
Nishiguchi	2003	[64, 68]	–	–	–	□	–	–	Type II errors	489	920	53.15%	**0.0301**
Hamasaki	2004	[62, 64]	–	–	–	?	–	?	Neutral	944	1848	51.08%	0.1821
Nishiguchi	2005	[58]	–	–	–	□	–	–	Type II errors	418	864	48.38%	0.8381
Repp	2006	[71]	-	□	□	□	□	-	Type II errors	42	86	48.84%	0.6267
Meyer	2007	[63]	□	–	?	□	□	□	Type II errors	276	554	49.82%	0.5507
Woszyck	2007	[69]	–	?	–	?	–	?	Type II errors	54	114	47.37%	0.7439
Pras	2010	[66]	?	–	?	?	–	?	Neutral	368	707	52.05%	0.1462
King	2012	[72]	–	□	–	?	□	?	Type II errors	34	61	55.74%	0.2213
KanetadaA	2013	[24]	?	–	?	–	–	–	Type I errors	62	108	57.41%	0.0743
KanetadaB	2013	[24]	?	–	?	–	–	–	Type I errors	135	224	60.27%	**1.281E-03**
Jackson	2014	[11, 65]	–	?	–	–	–	–	Neutral	585	960	60.94%	**6.352E-12**
Mizumachi	2015	[70]	?	–	?	–	–	–	Type I errors	86	136	63.24%	**1.279E-03**
Jackson	2016	[65]	–	?	–	–	–	–	Neutral	819	1440	56.88%	**1.000E-07**
									Total	**6736**	**12645**	**53.27%**	**1.006E-13**

may contribute towards Type II errors, i.e., an inability to demonstrate discrimination of high resolution audio.

Although full details of their experiment, methodology, and data are not available, some interesting secondary analysis is possible. [76] noted that "the percentage of subjects who correctly identified SACD at least 70% of the time appears to be implausibly low." In trials with at least 55 subjects, only one subject had 8 out of 10 correct and 2 subjects achieved 7 out of 10 correct. The probability of no more than 3 people getting at least 7 out of 10 correct by chance, is 0.97%. This suggests that the results were far from the binomial distribution that one would expect if the results were truly random.

If no one was able to distinguish between formats and there were no issues in the experimental design, then all trial results would be independent, regardless of whether the trials were by the same participant, and regardless of how participants are categorized. But [63] also gave a breakdown of correct answers by gender, age, audio experience, and hearing ability, depicted in Table 3. Non-audiophiles, in particular, have very low success rates, 30 out of 87, which has a probability of only ($p(X <= 30) = 0.25\%$). Chi squared analysis comparing audiophiles with non-audiophiles gives a p value of 0.18%, suggesting that it is extremely unlikely that the data for these two groups are independent. Similarly, analysis suggests that the results for those with and

Table 3. Statistical analysis of data from [63]. Statistically significant results at $\alpha = 0.05$ are given in bold.

Group		Correct	Incorrect	Total	p value	χ^2 statistic	p value-independence
Total trials		276	278	554	$p(X \geq 276) = 0.5507$	-	-
Gender	Male	258	248	506	$p(X \geq 258) = 0.3446$	3.1904	0.0741
	Female	18	30	48	$p(X \leq 18) = 0.0557$		
Hearing/Age	>15 kHz/14–25 years old	116	140	256	$p(X \leq 116) = 0.0752$	3.867	**0.0492**
	≤15 kHz/26 or more years old	160	138	298	$p(X \geq 160) = 0.1119$		
Experience	Audiophile/audio professional	246	221	467	$p(X \geq 246) = 0.1334$	9.7105	**0.0018**
	Non-audiophile	30	57	87	$p(X \leq 30) = \mathbf{0.0025}$		

without strong high frequency hearing also do not appear independent, $p = 4.92\%$. Note, however, that if there was a measurable effect, one would expect some dependency between answers from the same participant. The analysis in Table 3 is based only on total correct answers, not correct answers per participant, since this data was not available.

2.3 Multiple Comparisons

Some p value analysis was misleading. The discrimination tests all have a finite number of trials, each with dichotomous outcomes. Thus, they each give results with discrete probabilities, which may not align well with a given level of significance. For instance, if a discrimination trial is repeated 10 times with a participant, and $\alpha = 0.05$, then only 9 or 10 correct could give $p \leq \alpha$, even though this occurs by chance with probability $p = 1.07\%$, which is much less than the significance level. This low statistical power implies that a lack of participants with $p \leq \alpha$ may be less of an indicator of an inability to discriminate than it first appears. This should also be taken into consideration when accounting for multiple comparisons.

In several studies, a small number of participants had some form of evaluation with a p value less than 0.05. This is not necessarily evidence of high resolution audio discrimination, since the more times an experiment is run, the higher the likelihood that any result may appear significant by chance. Several experiments also involved testing several distinct hypotheses, e.g., does high resolution audio sound sharper, does it sound more tense, etc. Given enough hypotheses, some are bound to have statistical significance.

This well-known multiple comparisons problem was accounted for using the Holm, Holm-Bonferroni, and Sidak corrections (see Appendix), which all gave similar results, and we also looked at the likelihood of finding a lack of statistically significant results where no or very few low p values were found. This is summarized in Table 4, which also gives the actual significance levels given that each participant has a limited number of trials with dichotomous outcomes. Interestingly, the results in Table 4 agree with the results of retesting statistically significant individuals in Nishiguchi 2003 and Hamasaki 2004, confirm the statistical significance of several results in Yoshikawa 1995, and highlight the implausible lack of seemingly significant results among the test subjects in Meyer 2007, previously noted by [76]. For Pras 2010, they refute the significance of the specific individuals who "anti-discriminate" (consistently misidentify the high resolution content in an ABX test), but confirms the significance of there being 3 such individuals out of 16, and similarly for the 3 significant results out of 15 stimuli.

2.4 Hypotheses and Disputed Results

Many study results have been disputed, or given very different interpretations by their authors. Oohashi 1991 noted a persistence effect; when full range content (including frequencies beyond 20 kHz) is played immediately before low pass filtered content, the subjects incorrectly identified them as the same. Woszcyk 2007 found statistical significance in the different test conditions that were used and speculated that the complex high resolution signals might have been negatively perceived as artifacts. Both Oohashi 1991 and Woszcyk 2007 may have suffered a form of Simpson's paradox, where these false negatives canceled out a statistically significant discrimination of high resolution audio in other cases. Similar problems may have plagued King 2012, where many participants rated the "live feed" as sounding least close to live. Indeed, Pras 2010 observed a group of individuals who "anti-discriminate" and consistently misidentify high resolution audio in ABX tests.

Several studies intentionally considered discrimination of a high resolution format even if the content was not intended to be high resolution. In [62, 64], it was claimed that Nishiguchi 2003 did not have sufficient high frequency content. In one condition for Woszcyk 2007, a 20 kHz cut-off filter was used, and in Nishiguchi 2005 the authors stated that they "used ordinary professional recording microphones and did not intend to extend the frequency range intentionally during the recording sessions . . . sound stimuli were originally recorded using conventional recording microphones." These studies were still considered in the meta-analysis of Sec. 3 since further investigation (e.g., spectrograms and frequency response curves in [58, 64, 68]) shows that they may still have contained high frequency content, and the extent to which one can discriminate a high sample rate format without high frequency content is still a valid question.

Other studies noted conditions that may contribute to high resolution audio discrimination. [25, 60, 61] noted that intermodulation distortion may result in aliasing of high frequency content, and [63] remarked on the audibility of the noise floor for 16 bit formats at high listening levels. [23] had participants blindfolded, in order to eliminate visual distractions, and [56], though finding a null result when comparing two high resolution formats, still noted that the strongest results were among participants who conducted the test with headphones.

Together, the observations mentioned in this section provide insight into potential biases or flaws to be assessed for each study, and a set of hypotheses to be validated, if possible, in the following meta-analysis section.

2.5 Risk of Bias

Table 2 B presents a summary of the risk of bias, or other issues, in the studies. This has been adapted from [77], with a focus on the types of biases common to these tests. In particular, we are concerned with biases that may be introduced due to the methodology (e.g., the test may be biased towards inability to discriminate high resolution content if listeners are asked to select stimuli closest to "live" without defining "live," as in [72]), the experimental design (e.g., level imbalance as in [45, 46] or intermodulation distortion as in [25, 60, 61] may result in false positive discrimination), or the choice of stimuli (e.g., stimuli may not have contained high resolution content as in [58], or used test signals that may not capture whatever behavior might cause perception of high resolution content, as in [26, 59] leading to false

Table 4. Multiple comparisons testing. The last two rows tested the probability of a lack of significant results in multiple comparisons. The last column considers the probability of obtaining at least (or at most, for Nishiguchi 2005 and Meyer 2007) that many significant results given the significance level and number of tests.

Study	# tests	Repeated test type	Significance level	# significant	# significant corrected for multiple comparisons	p value
Yoshikawa 1995	22	Subject/stimuli	0.05	5	4	0.00402
Pras 2010	15	Stimuli	0.05	3	0	0.0362
Pras 2010	16	Subject	0.05 (2 sided)	3	0	0.0429
Nishiguchi 2003	36	Subject	0.0207; ≥15 of 20	1	0	0.5290
Hamasaki 2004	13	Subject	0.0411; ≥57 of 96	1	0	0.4204
Hamasaki 2004	39	Subject/stimuli	0.0251; ≥22 of 32	2	0	0.2556
Nishiguchi 2005	54	Subject/stimuli	0.0320; ≥17 of 24	0	–	0.1731
Meyer 2007	55	Subject	0.1719; ≥7 of 10	3	–	0.0097

negatives). We identified an unclear risk in each category if the risk had not been addressed or discussed and a high risk if there was strong evidence of a flaw or bias in a category. Potential biases led both to Type I and Type II errors, i.e., to falsely suggesting an ability to discriminate or not to discriminate high resolution content, though Type II errors were more common. Furthermore, biases often existed that might result in Type II errors even when the overall result demonstrated an effect (e.g., [59]).

3 META-ANALYSIS RESULTS

The most common way that results are presented in the studies are as the mean percentage of trials with correct discrimination of stimuli averaged over all participants. Thus this effect measure, equivalent to a mean difference [77], is used in most of the analysis that follows. The influence of these and other choices will be analyzed in Sec. 3.7.

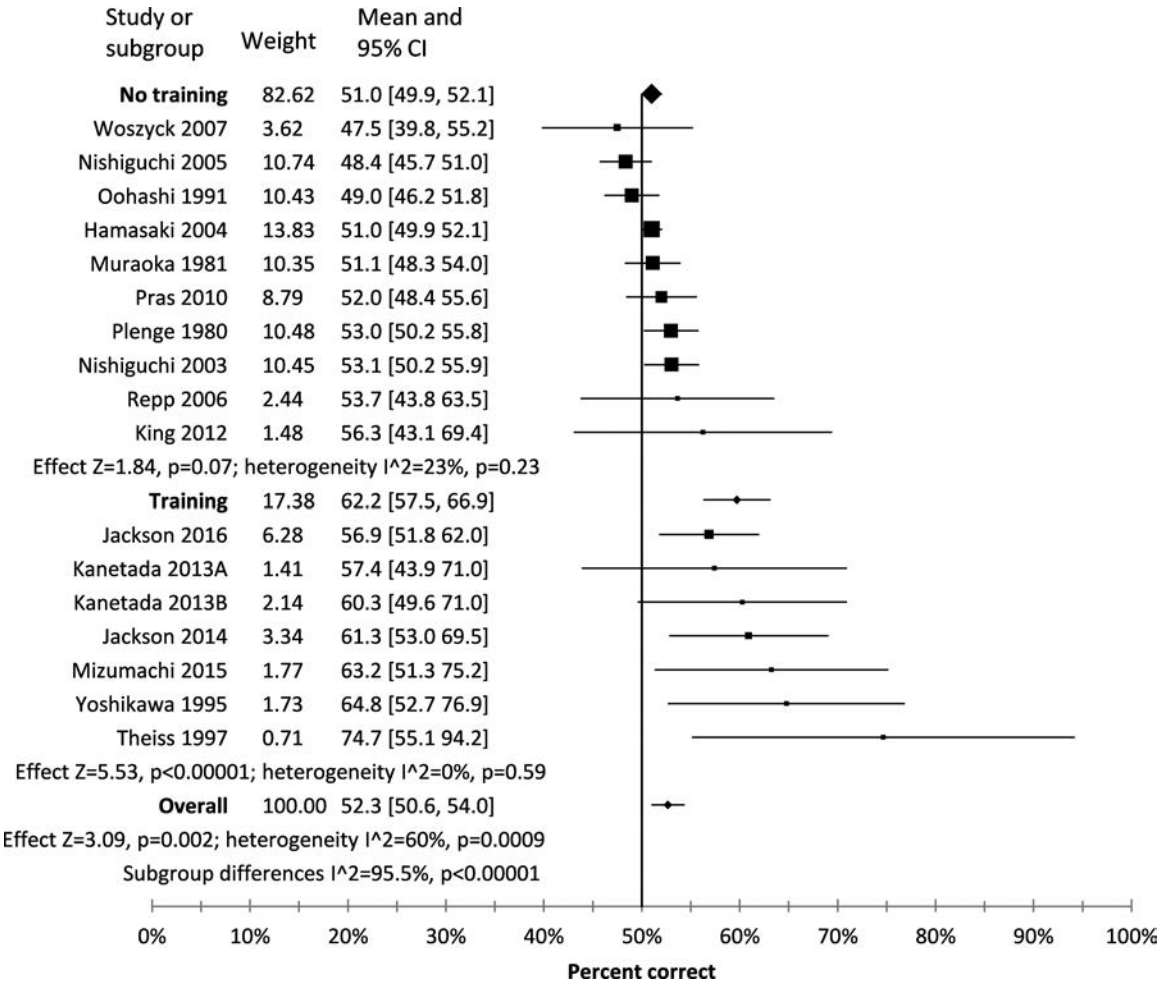

Fig. 2. A forest plot of studies where mean and standard deviation over all participants can be obtained, divided into subgroups based on whether participants were trained.

3.1 Binomial Tests

A simple form of analysis is to consider a null hypothesis, for each experiment, that there is no discernible effect. For all experimental methodologies, this would result in the answer for each trial, regardless of stimuli and subject, having a 50% probability of being correct. Table 2C depicts the number of trials, percentage of correct results for each trial, and the cumulative probability of at least that many correct answers if the experiment was truly random. Significant results at a level of $\alpha = 0.05$ are given in the last column of Table 2. Of note, several experiments where the authors concluded that there was not a statistically significant effect (Plenge 1980, Nishiguchi 2003), still appear to suggest that the null hypothesis can be rejected.

3.2 To What Extent Does Training Affect Results?

Fig. 2 depicts a forest plot of all studies where mean and standard deviation per participant can be obtained, divided into subgroups where participants either received detailed training (explanation of what to listen for, examples where artifacts could be heard, pretest with results provided to participants, etc.), or received no or minimal training (explanation of the interface, screening for prior experience in critical listening).

The statistic I^2 measures the extent of inconsistency among the studies' results and is interpreted as approximately the proportion of total variation in study estimates that is due to heterogeneity (differences in study design) rather than sampling error. Similarly, a low p value for heterogeneity suggests that the tests differ significantly, which may be due to bias.

The results are striking. The training subgroup reported an overall strong and significant ability to discriminate high resolution audio. Furthermore, tests for heterogeneity gave $I^2 = 0\%$ and $p = 0.59$, suggesting a strong consistency between those studies with training, and that all variation in study estimates could be attributed to sampling error. In contrast, those studies without training had an overall small effect. Heterogeneity tests reveal large differences between these studies $I^2 = 23\%$, though this may still be attributed to statistical variation, $p = 0.23$. Contrasting the subgroups, the test for subgroup differences gives $I^2 = 95.5\%$ and $p<10^{-5}$, suggesting that almost all variation in subgroup estimates is due to genuine variation across the "Training" and "No training" subgroups rather than sampling error.

3.3 How Does Duration of Stimuli and Intervals Affect Results?

The International Telecommunication Union recommends that sound samples used for sound quality comparison should not last longer than 15–20 s, and intervals between sound samples should be up to 1.5 s [78], partly because of limitations in short-term memory of test subjects. However, the extensive research into brain response to high resolution content suggests that exposure to high frequency content may evoke a response that is both lagged and persistent for tens of seconds, e.g., [22, 48]. This implies that effective testing of high resolution audio discrimination should use much longer samples and intervals than the ITU recommendation implies.

Unfortunately, statistical analysis of the effect of duration of stimuli and intervals is difficult. Of the 18 studies suitable for meta-analysis, only 12 provide information about sample duration and 6 provide information about interval duration, and many other factors may have affected the outcomes. In addition, many experiments allowed test subjects to listen for as long as they wished, thus making these estimates very rough approximations.

Nevertheless, strong results were reported in Theiss 1997, Kaneta 2013A, Kanetada 2013B and Mizumachi 2015, which all had long intervals between stimuli. In contrast, Muraoka 1981 and Pras 2010 had far weaker results with short duration stimuli. Furthermore, Hamasaki 2004 reported statistically significant stronger results when longer stimuli were used, even though participant and stimuli selection had more stringent criteria for the trials with shorter stimuli. This is highly suggestive that duration of stimuli and intervals may be an important factor.

A subgroup analysis was performed, dividing between those studies with stated long duration stimuli and/or long intervals (30 seconds or more) and those that state only short duration stimuli and/or short intervals. The Hamasaki 2004 experiment was divided into the two subgroups based on stimuli duration of either 85–120 s or approx. 20 s [62, 64].

The subgroup with long duration stimuli reported 57% correct discrimination, whereas the short duration subgroup reported a mean difference of 52%. Though the distinction between these two groups was far less strong than when considering training, the subgroup differences were still significant at a 95% level, $p = 0.04$. This subgroup test also has a small number of studies (14), and many studies in the long duration subgroup also involved training, so one can only say that it is suggestive that long durations for stimuli and intervals may be preferred for discrimination.

3.4 Effect of Test Methodology

There is considerable debate regarding preferred methodologies for high resolution audio perceptual evaluation. Authors have noted that ABX tests have a high cognitive load [11], which might lead to false negatives (Type II errors). An alternative, 1IFC Same-different tasks, was used in many tests. In these situations, subjects are presented with a pair of stimuli on each trial, with half the trials containing a pair that is the same and the other half with a pair that is different. Subjects must decide whether the pair represents the same or different stimuli. This test is known to be "particularly prone to the effects of bias [79]." A test subject may have a tendency towards one answer, and this tendency may even be prevalent among subjects. In particular, a subtle difference may be perceived but still identified as 'same," biasing this approach towards false negatives as well.

We performed subgroup tests to evaluate whether there are significant differences between those studies where subjects performed a 1 interval forced choice "same/different"

test, and those where subjects had to choose among two alternatives (ABX, AXY, or XY "preference" or "quality"). For same/different tests, heterogeneity test gave $I^2 = 67\%$ and $p = 0.003$, whereas $I^2 = 43\%$ and $p = 0.08$ for ABX and variants, thus suggesting that both subgroups contain diverse sets of studies (note that this test has low power, and so more importance is given to the I^2 value than the p value, and typically, α is set to 0.1 [77]).

A slightly higher overall effect was found for ABX, 0.05 compared to 0.02, but with confidence intervals overlapping those of the 1IFC "same/different" subgroup. If methodology has an effect, it is likely overshadowed by other differences between studies.

3.5 Effect of Quantization

Most of the discrimination studies focus on the effect of sample rate and the use of stimuli with and without high frequency content. It is well-known that the dynamic range of human hearing (when measured over a wide range of frequencies and considering deviations among subjects) may exceed 100 dB. Therefore, it is reasonable to speculate that bit depth beyond 16 bits may be perceived.

Only a small number of studies considered perception of high resolution quantization (beyond 16 bits per sample). Theiss 1997 reported 94.1% discrimination for one test subject comparing 96 kHz 24 bit to 48 kHz 16 bit, and the significantly lower 64.9% discrimination over two subjects comparing 96 kHz 16 bit to 48 kHz 16 bit. Jackson 2014 compared 192 kHz to 44.1 kHz and to 48 kHz with different quantizers. They found no effect of 24 to 16 bit reduction in addition to the change in sample rate. Kanetada 2013A, Kanetada 2013B, and Mizumachi 2015 all found strong results when comparing 16 to 24 bit quantization. Notably, Kanetada 2013B used 48 kHz sample rate for all stimuli and thus focused only on difference in quantization.

However, Kanetada 2013A, Kanetada 2013B, and Mizumachi 2015 all used undithered quantization. Dithered quantization is almost universally preferred since, although it increases the noise floor, it reduces noise modulation and distortion. But few have looked at perception of dither. [80] dealt solely with perception of the less commonly used subtractive dither, and only at low bit depths, up to 6 bits per sample. [81] investigated preference for dither for 4 to 12 bit quantizers in two bit increments. Interestingly, they found that at 10 or 12 bits, for all stimuli, test subjects either did not show a significant preference or preferred undithered quantization over rectangular dither and triangular dither for both subtractive and nonsubtractive dither. Jackson 2014 found very little difference (over all subjects and stimuli) in discrimination ability when dither was or was not applied. Thus, based on the evidence available, it is reasonable to include these as valid discrimination experiments even though dither was not applied.

3.6 Is there Publication Bias?

A common concern in meta-analysis is that smaller studies reporting negative or null results may not be published. To investigate potential publication bias, we produced a funnel plot of the 16 studies where a mean difference per participant was obtained, and funnel plots of the two subgroups of studies with and without training, Fig. 3. The overall funnel plot shows clear asymmetry, with few studies showing a low mean difference and a high standard error, i.e., few small studies with null results. Several studies also fall outside the 95% confidence interval, further suggesting biases. However, much of the asymmetry disappears when different funnel plots are provided for subgroups with and without training, and all studies fall within their confidence intervals. Though publication bias may still be a factor, it is likely that the additional effort in conducting a study with training was compensated for by less participants or less trials per participant, which contributes to larger standard errors. This is in full agreement with the cautions described in [82, 83].

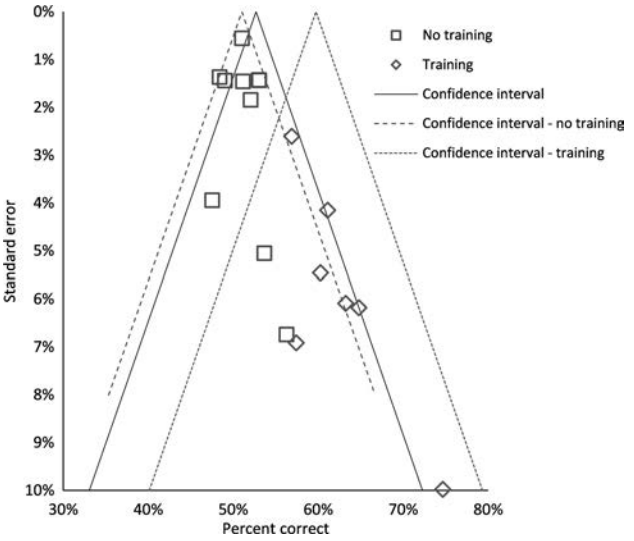

Fig. 3. Funnel plots of the 17 studies where a mean difference per participant was obtained, along with associated 95% confidence intervals. Much of the asymmetry in the overall funnel plot that might be attributed to publication bias is removed when funnel plots are given for subgroups of studies with and without training.

3.7 Sensitivity Analysis

This meta-analysis involves various decisions that may be considered subjective or even arbitrary. Most notably, we aimed to include all data from all high resolution perception studies that may be transformed into an average ratio, over all participants, of correct to total discrimination tasks. The choice of included studies, interpretation of data from those studies, and statistical approaches may all be questioned. For this reason, Table 5 presents a sensitivity analysis, repeating our analysis and subjecting our conclusions to alternative approaches.

Though the studies are diverse in their approaches, we considered fixed effect models in addition to random effect models. These give diminished (but still significant) results, primarily because large studies without training are weighed highly under such models.

We also considered treating the studies as yielding dichotomous rather than continuous results. That is, rather

Table 5. Sensitivity analysis showing the percent correct (effect estimate) and confidence intervals under different approaches to the meta-analysis. Data type was considered as either continuous (CONT) for means and standard errors over all participants, or dichotomous (DIC) for number of correct responses out of all trials. Inverse Variance (IV) and Mantel-Haenszel (MH) statistical methods were considered.

Name	Data Type	Statistical Method	# studies	Analysis Model	Effect Estimate	Confidence interval
All participants	CONT	IV	17	Random	52.7%	[51.0 54.4]
				Fixed	51.4%	[50.7 52.2]
All trials	DIC	IV	18	Random	54.4%	[52.1 56.6]
		MH			54.4%	[52.1 56.6]
		IV		Fixed	53.4%	[52.5 54.2]
		MH			53.3%	[52.4 54.1]
Conditions as separate studies	DIC	IV	32	Random	54.8%	[53.0 56.7]
Shared authors as single study	DIC	IV	13		54.1%	[51.5 56.7]
Sample rate only	CONT	IV	14		52.5%	[50.8 54.3]
Modern digital formats only	CONT	IV	12		52.9%	[50.8 55.0]

than mean and standard error over all participants, we simply consider the number of correctly discriminated trials out of all trials. This approach usually requires an experimental and control group, but due to the nature of the task and the hypothesis, it is clear that the control is random guessing, i.e., 50% correct as number of trials approaches infinity. This knowledge of the expected behavior of the control group allows use of standard meta-analysis approaches for dichotomous outcomes. Treating the data as dichotomous gave stronger results, even though it allowed inclusion of Meyer 2007, which was one of the studies that most strongly supported the null hypothesis. Use of the Mantel-Haenszel (as opposed to Inverse Variance) meta-analysis approach with the dichotomous data had no influence on results.

A full description of the statistical methods used for continuous and dichotomous results, fixed effects and random effects, and the Inverse Variance and Mantel-Haenszel methods, is given in the Appendix.

Many studies involved several conditions, and some authors participated in several studies. Treating each condition as a different study (a valid option since some conditions had quite different stimuli or experimental set-ups) or merging studies with shared authors was performed for dichotomous data only, since it was no longer possible to associate results with unique participants. Treating all conditions as separate studies yielded the strongest outcome. This is partly because some studies had conditions giving opposite results, thus hiding strong results when the different conditions were aggregated. Finally, we considered focusing only on sample rate and bandwidth (removing those studies that involved changes in bit depth) or only those using modern digital formats (removing the pre2000s studies that used either analogue or DAT systems). Though this excluded some of the studies with the strongest results, it did not change the overall effect.

Though not shown in Table 5, all of the conditions tested gave an overall effect with $p<0.01$, and all showed far stronger ability to discriminate high resolution audio when the studies involved training.

4 CONCLUSIONS

4.1 Implications for Practice

The meta-analysis herein was focused on discrimination studies concerning high resolution audio. Overall, there was a small but statistically significant ability to discriminate between standard quality audio (44.1 or 48 kHz, 16 bit) and high resolution audio (beyond standard quality). When subjects were trained, the ability to discriminate was far more significant. The analysis also suggested that careful selection of stimuli, including their duration, may play an important role in the ability to discriminate between high resolution and standard resolution audio. Sensitivity analysis, where different selection criteria and different analysis approaches were applied, confirmed these results. Potential biases in the studies leaned towards Type II errors, suggesting that the ability to discriminate high resolution audio may possibly be stronger than the statistical analysis indicates.

Several important practical aspects of high resolution audio perception could neither be confirmed nor denied. Most studies focused on the sample rate, so the ability to discriminate high bit depth, e.g., 24 bit versus 16 bit, remains an open question. None of the studies subjected to meta-analysis used headphones, so questions regarding how presentation over headphones affects perception also remain open. The meta-analysis also did not pursue questions regarding specific implementations of audio systems, such as the choice of filtering applied, the specific high resolution audio format that was chosen, or the influence of the various hardware components in the audio recording and reproduction chain (other than assessing potential biases that might be introduced by poor choices).

In summary, these results imply that, though the effect is perhaps small and difficult to detect, the *perceived fidelity of an audio recording and playback chain is affected by operating beyond conventional consumer oriented levels*. Furthermore, though the causes are still unknown, this perceived effect can be confirmed with a variety of statistical approaches and it can be greatly improved through training.

4.2 Implications for Experimental Design

Evaluation of high resolution audio discrimination involves testing the limits of perception, and it is clear from the presented meta-analysis that it is difficult to detect. It is thus important that good test procedures are carefully followed. In addition, the work herein suggests several recommendations for future experimental design in this field:

1. Training—Test subjects should be trained in how to discriminate, given examples and informed of their results in practice sessions before the test.
2. Experimental design—There are several issues in the experimental set-up that may lead to Type I or Type II errors. In all stages, the recording and playback system for high resolution audio needs to have sufficient bandwidth to reproduce the full range of frequency content. There should be no level imbalance or differences in processing between the signal paths for high resolution and normal resolution content. Distortion levels and dynamic range should be measured, and tweeters (if used) should be aimed at the listener. Where possible, this should be confirmed by measuring the end-to-end response of the playback system. In general, any potential artifacts, confounding factors or additional variables should be measured and accounted for.
3. Stimuli—The study authors should ensure that the stimuli contain high resolution content. Ideally, the signal received at the listener position should be measured to ensure that this is the case. Since little has been established about the causes of high resolution perception, a wide range of stimuli should be considered. Test signals should be used with care since they may lack whatever features are needed for perception. Also, long duration stimuli are preferred, with (where this is an option for the methodology) a sufficient interval between stimuli.
4. Methodology—In several studies, test subjects may have had multiple interpretations of the research question. Preference or quality questions may be clouded by the participants' prior assumptions, leading to Type II errors. The task given to subjects should be unambiguous, and all participants should have a similar understanding of that task.
5. Analysis—Analysis methods should be established prior to the experiment, and any post-hoc approaches should be clearly identified. An over-reliance on individual p values should be avoided, especially when there are a finite number of trials with dichotomous outcomes. Where possible, multiple comparisons should be corrected.
6. Reporting—A full description of the experimental set-up should be provided, including data sheets of the used equipment. The listening level at the listener position should be provided. Full data should be made available, including each participant's answers, the stimuli, and their presentation (duration, ordering) in each trial.

4.3 Implications for Meta-Analysis

The work presented herein is one of a very few, if any, papers that have applied rigorous and formal meta-analysis techniques to studies in the field of perceptual audio evaluation, or more generally, psychophysics. It has shown that techniques designed for studies involving intervention and control groups can be applied to experiments involving repeated trials with dichotomous outcomes, typically lacking a control. Measures of risk difference or mean difference, and their standard errors, can be adapted to situations where the mean value of the control (in this case, correct discrimination by pure guessing) is determined by probability theory.

This paper also uncovered interesting phenomena that needed to be considered in the analysis. Several studies, such as Oohashi 1991 and King 2012, showed evidence of Simpson's paradox, where opposite trends in the data may have led to little effect being observed. Others (Nishiguchi 2003 and Hamasaki 2004) may have employed an equivalent of the Martingale betting system, where an experiment was repeated with a participant until a lack of effect was observed (though this may also be considered a method of verifying an initial observation). And several studies had conclusions that may have suffered from the multiple comparisons problem (Yoshikawa 1995, Nishiguchi 2003, Hamasaki 2004, Pras 2010). Interestingly, several studies reported results suggesting that for some trials, participants had an uncanny ability to discriminate far worse than guessing (Oohashi 1991, Meyer 2007, Woszcyk 2007, Pras 2010).

We also uncovered an issue with the use of standard statistical hypothesis testing applied to multiple trials with dichotomous outcomes. This issue, which occurred in many studies, may lead to Type II errors, and to our knowledge has not been widely addressed elsewhere in the literature.

4.4 Future Research Directions

As previously mentioned, many proposed causes or factors in perception of high resolution audio could not be confirmed nor denied and warrant further investigation. Some of these questions are particularly intriguing, such as differences in perception over headphones versus loudspeakers, the effect of spatial audio rendering, the effect of quantization, the effect of duration (e.g., the trade-off between short-term auditory memory and the persistent effect of exposure to high frequency content), and the identification of critical stimuli where differences between high and standard resolution are most easily perceived.

There is a strong need for several listening tests. First, it is important that all test results be published. Notably, there is still a potential for reporting bias. That is, smaller studies that did not show an ability to discriminate high resolution content may not have been published. Second, it would be interesting to perform a subjective evaluation incorporating all of the design choices that, while not yielding Type I errors, were taken in those studies with the strongest discrimination results, e.g., Theiss 1997 had test subjects blindfolded to eliminate any visual distraction. If these procedures are followed, one might find that the ability to discriminate high resolution content is even higher than any reported study. Finally, no research group has mirrored the test design of another team, so there is need for an experiment that would provide independent verification of some of the more high profile or interesting reported results.

Many studies, reviewed in Sec. 1, involved indirect discrimination of high resolution audio, or focused on the limits of perceptual resolution. These studies were not included in the meta-analysis in order to limit our investigation to those studies focused on related questions of high interest, and amenable to systematic analysis. Further analysis should consider these additional listening tests. Such tests might offer insight both on causes of high resolution audio perception and on good test design, and might allow us to provide stronger results in some aspects of the meta-analysis.

However, many of these additional studies resulted in data that do not fit any of the standard forms for meta-analysis. Research is required for the development of statistical techniques that either transform the data into a more standard form, or establish a means of meta-analysis based on the acquired data. Finally, further research into statistical hypothesis testing of (multiple comparisons of) multiple trials with dichotomous outcomes would be useful for interpreting the results of many studies described herein and widely applicable to other research.

Additional data and analysis is available from code.soundsoftware.ac.uk/projects/hi-res-meta-analysis.

5 ACKNOWLEDGEMENTS

The author would like to express the deepest gratitude to all authors who provided additional data or insights regarding their experiments, including Helen Jackson, Michael Capp, Bob Stuart, Mitsunori Mizumachi, Amandine Pras, Brad Meyer, David Moran, Brett Leonard, Richard King, Wieslaw Woszczyk and Richard Repp. The author is also very grateful for the advice and support from Vicki Melchior, George Massenburg, Bob Katz, Bob Schulein, and Juan Adriano.

6 REFERENCES

[1] A. Flexer, et al., "A MIREX Meta-Analysis of Hubness in Audio Music Similarity," in *Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Porto, Portugal (2012).

[2] J. B. L. Smith and E. Chew, "A Meta-Analysis of the MIREX Structure Segmentation Task," in *Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Curitiba, Brazil (2013).

[3] M. McVicar, et al., "Automatic Chord Estimation from Audio: A Review of the State of the Art, " *IEEE/ACM Trans. Audio, Speech Lang. Proc.*, vol. 22 (2014). doi:10.1109/TASLP.2013.2294580

[4] E. Hemery and J.-J. Aucouturier, "One Hundred Ways to Process Time, Frequency, Rate and Scale in the Central Auditory System: A Pattern-Recognition Meta-Analysis," *Frontiers in Computational Neuroscience*, 03 July 2015. doi:10.3389/fncom.2015.00080

[5] R. J. Wilson, "Special Issue Introduction: High-Resolution Audio," *J. Audio Eng. Soc.*, vol. 52, p. 116 (2004 Mar.).

[6] J. R. Stuart and P. G. Craven, "A Hierarchical Approach to Archiving and Distribution," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9178.

[7] H. van Maanen, "Requirements for Loudspeakers and Headphones in the 'High Resolution Audio' Era," presented at the *AES 51st International Conference: Loudspeakers and Headphones* (2013 Aug.), conference paper 1-3

[8] J. R. Stuart, "Coding for High-Resolution Audio Systems," *J. Audio Eng. Soc.*, vol. 52, pp. 117–144 (2004 Mar.)

[9] J. R. Stuart, "High-Resolution Audio: A Perspective," *J. Audio Eng. Soc.*, vol. 63, pp. 831–832 (2015 Oct.).

[10] W. Woszczyk, "Physical and Perceptual Considerations for High-Resolution Audio," presented at the *115th Convention of the Audio Engineering Society* (2003 Oct.), convention paper 5931.

[11] H. M. Jackson, et al., "The Audibility of Typical Digital Audio Filters in a High-Fidelity Playback System," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9174.

[12] J. Vanderkooy, "A Digital-Domain Listening Test for High-Resolution," presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), convention paper 8203.

[13] B. Smagowska and M. Pawlaczyk-Łuszczyńska, "Effects of Ultrasonic Noise on the Human Body—A Bibliographic Review," *Int. J. Occupational Safety and Ergonomics (JOSE)*, vol. 19, pp. 195–202 (2013). doi:10.1080/10803548.2013.11076978

[14] W. B. Snow, "Audible Frequency Ranges of Music, Speech, and Noise," *J. Acoust. Soc. Am.*, vol. 3, pp. 155–166 (1931). doi:10.1121/1.1915552

[15] D. Gannett and I. Kerney, "The Discernibility of Changes in Program Band Width," *Bell Systems Technical J.*, vol. 23, pp. 1–10 (1944). doi:10.1002/j.1538-7305.1944.tb03144.x

[16] H. Fletcher, *Speech and Hearing in Communication* (Princeton, New Jersey: Van Nostrand, 1953).

[17] T. Zislis and J. L. Fletcher, "Relation of High Frequency Thresholds to Age and Sex," *J. Aud. Res.*, vol. 6, pp. 189–198 (1966)

[18] J. D. Harris and C. K. Meyers, "Tentative Audiometric Threshold Level Standards from 8 to 18 kHz,"

J. Acoust. Soc. Am., vol. 49, pp. 600–608 (1971). doi:10.1121/1.1912392

[19] J. L. Northern, et al., "Recommended High-Frequency Audiometric Threshold Levels (8000–18000 Hz)," *J. Acoust. Soc. Am.*, vol. 52, pp. 585–595 (1972). doi:10.1121/1.1913149

[20] D. R. Cunningham and C. P. Goetzinger, "Extra-High Frequency Hearing Loss and Hyperlipidemia," *Audiology*, vol. 13, pp. 470–484 (1974). doi:10.3109/00206097409071710

[21] S. A. Fausti, et al., "A System for Evaluating Auditory Function from 8000–20000 Hz," *J. Acoust. Soc. Am.*, vol. 66, pp. 1713– (1979).

[22] T. Oohashi, et al., "Multidisciplinary Study on the Hypersonic Effect," in *Interareal Coupling of Human Brain Function,Int. Congress Series* vol. 1226: (Elsevier, 2002), pp. 27–42. doi:10.1016/s0531-5131(01)00494-0

[23] B. Theiss and M. O. J. Hawksford, "Phantom Source Perception in 24 Bit @ 96 kHz Digital Audio," presented at the *103rd Convention of the Audio Engineering Society* (1997 Sep.), convention paper 4561.

[24] N. Kanetada, et al., "Evaluation of Sound Quality of High Resolution Audio," in *1st IEEE/IIAE Int. Conf. Intelligent Systems and Image Processing* (2013). doi:10.12792/icisip2013.014

[25] K. Ashihara and S. Kiryu, "Audibility of Components above 22 kHz in a Harmonic Complex Tones," *Acustica - Acta Acustica*, vol. 89, pp. 540–546 (2003)

[26] W. Bulla, "Detection of High-Frequency Harmonics in a Complex Tone," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), convention paper 4561.

[27] J.-E. Mortberg, "Is Dithered Truncation Preferred over Pure Truncation at a Bit Depth of 16 Bits when a Digital Requantization has Been Performed on a 24 Bit Sound File?," Bachelor, Lulea University of Technology (2007).

[28] M. L. Lenhardt, et al., "Human Ultrasonic Speech Perception," *Science*, vol. 253, pp. 82–85 (1991). doi:10.1126/science.2063208

[29] S. Nakagawa and S. Kawamura, "Temporary Threshold Shift in Audition Induced by Exposure to Ultrasound via Bone Conduction," in *27th Annual Meeting Int. Soc. Psychophysics*, Herzliya, Israel (2011).

[30] T. Hotehama and S. Nakagawaa, "Modulation Detection for Amplitude-Modulated Bone-Conducted Sounds with Sinusoidal Carriers in the High- and Ultrasonic-Frequency Range," *J. Acoust. Soc. Am.*, vol. 128, pp. 3011 (2010 Nov.). doi:10.1121/1.3493421

[31] K. Krumbholz, et al., "Microsecond Temporal Resolution in Monaural Hearing without Spectral Cues?" *J. Acoust. Soc. Am.*, vol. 113, pp. 2790–2800 (2003). doi:10.1121/1.1547438

[32] M. N. Kunchur, "Audibility of Temporal Smearing and Time Misalignment of Acoustic Signals," *Technical Acoustics*, vol. 17 (2007)

[33] M. N. Kunchur, "Temporal Resolution of Hearing Probed by Bandwidth Restriction," *Acta Acustica united with Acustica*, vol. 94, pp. 594–603 (2008). doi:10.3813/AAA.918069

[34] M. Kunchur, "Probing the Temporal Resolution and Bandwidth of Human Hearing," in *Proceedings of Meetings on Acoustics*, New Orleans (2007). doi:10.1121/1.2998548

[35] T. Muraoka, et al., "Examination of Audio Bandwidth Requirements for Optimum Sound Signal Transmission," *J. Audio Eng. Soc.*, vol. 29, pp. 2–9 (1981 Jan./Feb.).

[36] K. Ashihara, et al., "Hearing Thresholds in Free-Field for Pure Tone above 20 kHz," in *Int. Cong. Acoustics (ICA)* (2004).

[37] K. Ashihara, et al., "Hearing Threshold for Pure Tones above 20 kHz," *Acoust. Sci. & Technology*, vol. 27, pp. 12–19 (2006 Jan.).

[38] K. Ashihara, "Hearing Thresholds for Pure Tones above 16 kHz," *JASA Express Letters* (2007 Aug.).

[39] M. Omata, et al., "A Psychoacoustic Measurement and ABR for the Sound Signals in the Frequency Range between 10 kHz and 24 kHz," presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), convention paper 7566.

[40] M. Koubori, et al., "Psychoacoustic Measurement and Auditory Brainstem Response in the Frequency Range between 10 kHz and 30 kHz," presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), convention paper 8294.

[41] J. N. Oppenheim and M. O. Magnasco, "Human Time-Frequency Acuity Beats the Fourier Uncertainty Principle," *Physical Review Letters*, vol. 110 (Jan. 25 2013). doi:10.1103/PhysRevLett.110.044301

[42] M. Majka, et al., "Hearing Overcome Uncertainty Relation and Measure Duration of Ultrashort Pulses," *Europhysics News*, vol. 46, pp. 27–31 (2015). doi:10.1051/epn/2015105

[43] T. Oohashi, et al., "High-Frequency Sound Above the Audible Range Affects Brain Electric Activity and Sound Perception," presented at the *91st Convention of the Audio Engineering Society* (1991 Oct.), convention paper 3207.

[44] T. Oohashi, et al., "Inaudible High-Frequency Sounds Affect Brain Activity: Hypersonic Effect," *J. Neurophysiology*, vol. 83, pp. 3548–3558 (2000).

[45] R. Yagi, et al., "Auditory Display for Deep Brain Activation: Hypersonic Effect," in *Int. Conf. Auditory Display*, Kyoto, Japan (2002)

[46] R. Yagi, et al., "Modulatory Effect of Inaudible High-Frequency Sounds on Human Acoustic Perception," *Neuroscience Letters*, vol. 351, pp. 191–195 (2003). doi:10.1016/j.neulet.2003.07.020

[47] A. Fukushima, et al., "Frequencies of Inaudible High-Frequency Sounds Differentially Affect Brain Activity: Positive and Negative Hypersonic Effects," *PLOS One*(2014). doi:10.1371/journal.pone.0095464

[48] R. Kuribayashi, et al., "High-Resolution Music with Inaudible High-Frequency Components Produces a Lagged Effect on Human Electroencephalographic Activities," *Clinical Neuroscience*, vol. 25, pp. 651–655 (2014 Jun.). doi:10.1097/wnr.0000000000000151

[49] S. Han-Moi, et al., "Inaudible High-Frequency Sound Affects Frontlobe Brain Activity," *Contemporary Engineering Sciences*, vol. 23, pp. 1189–1196 (2014).

[50] M. Honda, et al., "Functional Neuronal Network Subserving the Hypersonic Effect," in *Int. Cong. Acoustics (ICA)*, Kyoto (2004).

[51] T. Oohashi, et al., "The Role of Biological System other than Auditory Air-Conduction in the Emergence of the Hypersonic Effect," *Brain Research*, vol. 1073-1074, pp. 339–347 (2006). doi:10.1016/j.brainres.2005.12.096

[52] M. Higuchi, et al., "Ultrasound Influence on Impression Evaluation of Music," *IEEE Pacific Rim Conf. Comm., Comp. and Sig. Proc. (PacRim)* (2009 Aug.).

[53] M. Kamada and K. Toraichi, "Effects of Ultrasonic Components on Perceived Tone Quality," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing* (1989). doi:10.1109/ICASSP.1989.266850

[54] S. Yoshikawa, et al., "Does High Sampling Frequency Improve Perceptual Time-Axis Resolution of Digital Audio Signal?" presented at the *103rd Convention of the Audio Engineering Society* (1997 Sep.), convention paper 4562.

[55] R. Yagi, et al., "A Method for Behavioral Evaluation of the 'Hypersonic Effect,' " *Acoust. Sci. & Tech.*, vol. 24, pp. 197–200 (2003).

[56] D. Blech and M. Yang, "DVD-Audio versus SACD: Perceptual Discrimination of Digital Audio Coding Formats," presented at the *116th Convention of the Audio Engineering Society* (2004 May), convention paper 6086.

[57] A. Marui, et al., "Subjective Evaluation of High Resolution Recordings in PCM and DSD Audio Formats," presented at the *136th Convention of the Audio Engineering Society* (2014 Apr.), convention paper 9019.

[58] T. Nishiguchi and K. Hamasaki, "Differences of Hearing Impressions among Several High Sampling Digital Recording Formats," presented at the *118th Convention of the Audio Engineering Society* (2005 May), convention paper 6469.

[59] G. Plenge, et al., "Which Bandwidth Is Necessary for Optimal Sound Transmission?" *J. Audio Eng. Soc.*, vol. 28, pp. 114–119 (1980 Mar.).

[60] K. Ashihara, "Audibility of Complex Tones above 20 kHz," *29th Int. Cong. and Exhibition on Noise Control Engineering (InterNoise)* (2000)

[61] K. Ashihara and S. Kiryu, "Detection Threshold for Tones above 22 kHz," presented at the *110th Convention of the Audio Engineering Society* (2001 May), convention paper 5401.

[62] K. Hamasaki, et al., "Perceptual Discrimination of Very High Frequency Components in Musical Sound Recorded with a Newly Developed Wide Frequency Range Microphone," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), convention paper 6298.

[63] E. B. Meyer and D. R. Moran, "Audibility of a CD-Standard A/D/A Loop Inserted into High-Resolution Audio Playback," *J. Audio Eng. Soc.*, vol. 55, pp. 775–779 (2007 Sep.).

[64] T. Nishiguchi, et al., "Perceptual Discrimination of Very High Frequency Components in Wide Frequency Range Musical Sound," *Applied Acoustics*, vol. 70, pp. 921–934 (2009). doi:10.1016/j.apacoust.2009.01.002

[65] H. M. Jackson, et al., "Further Investigations of the Audibility of Digital Audio Filters in a High-Fidelity Playback System," *J. Audio Eng. Soc. (under review)*, 2016

[66] A. Pras and C. Guastavino, "Sampling Rate Discrimination: 44.1 kHz vs. 88.2 kHz," presented at the *128th Convention of the Audio Engineering Society* (2010 May), convention paper 8101.

[67] S. Yoshikawa, et al., "Sound Quality Evaluation of 96 kHz Sampling Digital Audio," presented at the *99th Convention of the Audio Engineering Society* (1995 Oct.), convention paper 4112.

[68] T. Nishiguchi, et al., "Perceptual Discrimination between Musical Sounds with and without Very High Frequency Components," presented at the *115th Convention of the Audio Engineering Society* (2003 Oct.), convention paper 5876.

[69] W. Woszczyk, et al., "Which of the Two Digital Audio Systems Best Matches the Quality of the Analog System?" presented at the *AES 31st Int. Conference: New Directions in High Resolution Audio* London, (2007 June), conference paper 30.

[70] M. Mizumachi, et al., "Subjective Evaluation of High Resolution Audio Under In-car Listening Environments," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9226.

[71] R. Repp, "Recording Quality Ratings by Music Professionals," in *Int. Comp. Music Conf. (ICMC)*, New Orleans (2006).

[72] R. King, et al., "How Can Sample Rates Be Properly Compared in Terms of Audio Quality?" presented at the *133rd Convention of the Audio Engineering Society* (2012 Oct.), eBrief 77.

[73] B. C. J. Moore, "The Role of Temporal Fine Structure Processing in Pitch Perception, Masking, and Speech Perception for Normal-Hearing and Hearing-Impaired People," *J. Assoc. Res. Otolaryngology*, vol. 9, pp. 399–406 (2008). doi:10.1007/s10162-008-0143-x

[74] P. G. Craven, "Antialias Filters and System Transient Response at High Sample Rates," *J. Audio Eng. Soc.*, vol. 52, pp. 216–242 (2004 Mar.).

[75] G. S. Thekkadath and M. Spanner, "Comment on 'Human Time-Frequency Acuity Beats the Fourier Uncertainty Principle,' " *Physical Review Letters*, vol. 114 (2015).

[76] D. Dranove, "Comments on 'Audibility of a CD-standard A/D/A Loop Inserted into High-Resolution Audio Playback,' " *J. Audio Eng. Soc.*, vol. 58, p. 173–174 (2010 Mar.).

[77] J. P. T. Higgins and S. Green, Eds., *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0* (The Cochrane Collaboration, 2011).

[78] ITU, "Subjective assessment of sound quality, Recommendation BS.562-3" (1990).

[79] F. A. A. Kingdom and N. Prins, *Psychophysics: A Practical Introduction* (Academic Press, 2009).

[80] L. R. Rabiner and J. A. Johnson, "Perceptual Evaluation of the Effects of Dither on Low Bit Rate PCM Systems," *The Bell System Technical J.*, vol. 51, (1972 Sep.). doi:10.1002/j.1538-7305.1972.tb02665.x

[81] P. Kvist, et al., "A Listening Test of Dither in Audio Systems," presented at the *118th Convention of the Audio Engineering Society* (2005 May), convention paper 6328.

[82] M. Egger, et al., "Bias in Meta-Analysis Detected by a Simple, Graphical Test," *BMJ* vol. 315, pp. 629–634 (1997). doi:10.1136/bmj.315.7109.629

[83] J. Lau, et al., "The Case of the Misleading Funnel Plot," *BMJ*, vol. 333, pp. 597–600 (2006). doi:10.1136/bmj.333.7568.597

[84] J. J. Deeks and J. P. T. Higgins, "Statistical Algorithms in Review Manager 5," Statistical Methods Group of the Cochrane Collaboration, August 2010

[85] N. Mantel and W. Haenszel, "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," *J. National Cancer Institute*, vol. 22, pp. 719–748 (1959).

[86] R. DerSimonian and N. Laird, "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials*, vol. 7, pp. 177–188 (1986). doi:10.1016/0197-2456(86)90046-2

APPENDIX. STATISTICAL METHODS

This paper addresses concerns in the audio engineering and psychoacoustics community utilizing a variety of techniques from the field of meta-analysis. As such, much of the approach, terminology, and statistical methods may not be known to the readers. The meta-analysis techniques used herein are all based on the protocols and guidance for preparation of intervention systematic reviews [77] set forth by the Cochrane Collaboration Group, a global independent network that includes experts on the methodology of systematic reviews and meta-analysis. In this appendix we explain the meta-analysis techniques and statistical methods that were used and how they were adapted to the types of studies that were investigated.

We consider a meta-study of N studies involving a total of T trials, of which C trials resulted in correct discrimination between high resolution and normal resolution audio. Study n has S_n subjects and a total of T_n trials, and subject s within study n performs $T_{n,s}$ trials. C, C_n, $C_{n,s}$ are analogous to T, T_n, $T_{n,s}$, and correspond to the total number of correct trials in the meta-study, the total number of correct trials in study n and the number of correct trials by subject s in study n, respectively.

The most common way that results of the perceptual studies herein were presented was as a percent correct over all trials (i.e., *continuous* outcomes) for each participant, which gives an overall effect E_n and standard error $\mathrm{SE}\{E_n\}$ for the study n as

$$\begin{aligned} E_n &= \textstyle\sum_{S_n} (C_{n,s}/T_{n,s})/S_n \\ SE\{E_s\} &= \sqrt{\frac{\sum_{S_n} (C_{n,s}/T_{n,s} - E_n)^2}{S_n(S_n-1)}}. \end{aligned} \tag{A.1}$$

Unlike most meta-analysis, an important aspect of these studies is that none of them involve a control group. However, all studies were constructed in such a way that, if the experimental design did not introduce biases, then the mean percent correct should be 50%. Thus this percent correct minus 50% is similar to a risk difference or difference of means in standard meta-analysis.

Alternatively, study outcomes may be represented as *dichotomous* results considering the overall number of correct results, giving the effect and it's standard error (as in [84], Bessel's correction was not used),

$$\begin{aligned} E_n &= C_n/T_n \\ SE\{E_n\} &= \sqrt{C_n(T_n - C_n)/{T_n}^3}. \end{aligned} \tag{A.2}$$

Multiple Comparisons Testing

In Sec. 2.3, multiple comparisons testing was performed on studies that reported statistically significant results from test subjects. For a given study we ordered the S_n subject results in terms of their corresponding p values, ordered from lowest to highest p.

For the Bonferroni method, the significance level α was replaced with α/S_n. For the Holm–Bonferroni method, only the first k subjects such that $p_i \leq \alpha/(S_n + 1 - i)$ for all subjects $i \leq k$, where α is the uncorrected significance level, are considered to have statistically significant results. For the Sidak correction, assuming non-negative dependence, $\alpha/(S_n + 1 - i)$ is replaced with $1 - (1 - \alpha)^{1/(S_n+1-i)}$ resulting in a slightly more powerful test. The same procedure was used for other multiple comparisons tests reported in Table 4 by replacing the number of subjects S_n with the equivalent number of stimuli or subject/stimuli pairs.

Fixed Effect Meta-Analysis

In the inverse variance method, individual effect sizes are weighted by the reciprocal of their variance,

$$w_n = 1/SE^2\{E_n\}, \tag{A.3}$$

from either (A.1) for continuous outcomes or (A.2) for dichotomous outcomes. Studies are combined to give a summary estimate and associated summary error,

$$\begin{aligned} E_{IV} &= \textstyle\sum w_n E_n / \sum w_n \\ \mathrm{SE}\{E_{IV}\} &= 1/\sqrt{\textstyle\sum w_n} \end{aligned} \tag{A.4}$$

If the studies are expressed as dichotomous results, Eq. (A.2), then the Mantel-Haenszel method may be used [85]. Each study is given weight T_n, so the summary effect and associated standard error are

$$\begin{aligned} E_{MH} &= \textstyle\sum T_n E_n / T \\ \mathrm{SE}\{E_{MH}\} &= \sqrt{\textstyle\sum \frac{C_n(T_n - C_n)}{T_n}}/T \end{aligned} \tag{A.5}$$

For both inverse variance and Mantel-Haenszel methods, the heterogeneity Q and I^2 statistics are given by:

$$\begin{aligned} Q &= \textstyle\sum w_n (E_n - E)^2 \\ I^2 &= (1 - (S-1)/Q) \times 100\% \end{aligned}, \tag{A.6}$$

where E is the summary estimate from either Eq. (A.4) or Eq. (A.5) and the w_n are the weights calculated from (A.3). I^2measures the extent of inconsistency among the studies' results, and is interpreted as approximately the proportion of total variation in study estimates that is not due to sampling error.

Random-Effects Meta-Analysis

Under the random-effects model [86], the assumption of a common intervention effect is relaxed, and the effect sizes are assumed to have a normal distribution with variance estimated as

$$\tau^2 = \max[\frac{Q-(N-1)}{\sum w_n - \sum w_n^2/\sum w_n}, 0], \tag{A.7}$$

where Q is from Eq. (A.6) using either the inverse invariant or Mantel-Haenszel method, and the w_n are the inverse-variance weights from Eq. (A.3). Each study is given weight

$$w'_n = 1/(SE\{E_n\}^2 + \tau^2) \tag{A.8}$$

The summary effect size and its standard error are given by

$$\begin{aligned} E &= \sum w'_n E_n / \sum w'_n \\ SE\{E\} &= 1/\sqrt{\sum w'_n} \end{aligned} \tag{A.9}$$

When Q is less than or equal to $N-1$, the estimate of the between study variation τ^2 is zero, and the weights coincide with those given by the inverse-variance method.

Meta-Analysis Test Statistics

The 95% confidence interval for E is given by $E \pm 1.96\text{SE}\{E\}$. The test statistic for presence of an overall intervention effect is given by $Z = E/\text{SE}\{E\}$. Under the null hypothesis that there is no overall effect this follows a standard normal distribution.

The test for comparison of subgroups is based on testing for heterogeneity across G subgroups rather than across N studies. Let E_g be the summary effect size for subgroup g, with standard error $\text{SE}\{E_g\}$. The summary effect size may be based on either a fixed-effect or a random-effects meta-analysis. These numbers correspond to Eq. (A.5) for Mantel-Haenszel or Eq. (A.4) for inverse-variance under fixed-effect meta-analyses, or Eq. (A.9) for random-effects meta-analyses, each applied within each subgroup. A weight for each subgroup is computed:

$$w_g = 1/\text{SE}^2\{E_g\}, \tag{A.10}$$

then a summary effect size across subgroups is found using a fixed-effect meta-analysis:

$$E_G = \sum w_g E_g / \sum w_g \tag{A.11}$$

Statistics for differences across subgroups are $Q_G = \sum w_g (E_g - E_G)^2$ and $I^2 = (1-(G-1)/Q_G) \times 100\%$.

THE AUTHOR

Josh Reiss

Josh Reiss is a Reader with Queen Mary University of London's Centre for Digital Music, where he leads the audio engineering research team. He received his Ph.D. from Georgia Tech, specializing in analysis of nonlinear systems. His early work on sigma delta modulators led to patents and an IEEE best paper award nomination. He has investigated music retrieval systems, time scaling and pitch shifting techniques, polyphonic music transcription, loudspeaker design, automatic mixing, sound synthesis, and digital audio effects. His primary focus of research, which ties together many of the above topics, is on the use of state-of-the-art signal processing techniques for professional sound engineering.

Dr. Reiss has published over 160 scientific papers, including more than 70 AES publications. His co-authored publication, "Loudness Measurement of Multitrack Audio Content Using Modifications of ITU-R BS.1770," was recipient of the 134th AES Convention's Best Peer-Reviewed Paper Award. He co-authored the textbook *Audio Effects: Theory, Implementation and Application*. He is co-founder of the start-up company LandR, providing intelligent tools for audio production. He is a former governor of the AES, and was Chair of the 128th, Papers Chair of the 130th, and Co-Papers Chair of the 138th AES Conventions.