The Placebo Method, A Comparison of In-Situ Subjective Evaluation Methods for Vehicles

5136 (L - 5)

Neal House and Roger Shively
Harman Motive Inc., Martinsville, USA

Presented at
the 108th Convention
2000 February 19-22
Paris, France

AES

AN AUDIO ENGINEERING SOCIETY PREPRINT

The Placebo Method, A Comparison of In-Situ Subjective Evaluation Methods for Vehicles

Neal House
Roger Shively

Harman-Motive Inc., Martinsville, Indiana, USA

The "placebo" subjective evaluation method is investigated for sighted in-situ testing of vehicle sound systems. The placebo method is designed to reduce or eliminate bias effects that may occur with typical sighted in-situ subjective evaluations. With the placebo method, some modified program selections are added to the normal program sets that the listeners are to evaluate. Program modifications can be made to one or more spectral, spatial, dynamic, or temporal parameters. The listeners are instructed that the modified programs are randomly distributed in their program list, and the vehicle's sound system, the modified programs, or both could effect the judgements they make. Hypothetically, this should cause the listeners to focus on judging each individual program selection instead of the vehicle. The placebo method is compared statistically and functionally against blind and sighted in-situ methods and the results are given.

0 BACKGROUND

Studies on consumer loudspeaker's [1,2] have shown that sound quality judgements can be significantly biased by non-auditory factors such as size, price and name brand. It has been shown that the effects of bias or prejudice on the listener's judgements can be more significant than the effects due to differences in the loudspeakers. Conducting subjective evaluations on vehicle sound systems are particularly challenging because of difficulty isolating the listener from the system/environment they are to evaluate [3,4]. Three subjective evaluation methods (i.e. sighted in-situ, blind in-situ and blind-binaural) have been used in the recent past to judge and rank qualities of automotive sound systems. Each of these methods has specific advantages and disadvantages regarding bias reduction, attribute resolution, and practical application.

The most commonly used evaluation method for vehicle sound systems is the sighted in-situ method. This is because it is the fastest, easiest and most natural means to collect information and compare sound systems. This method is basically an extension of the sound system engineer's normal routine in designing and evaluating their work. It is also the most natural in terms of how a person would sit in, listen and use a vehicle on a normal basis, which includes both static (0 mph) and dynamic (at velocity) evaluations under realistic conditions. Unfortunately, this method cannot remove the non-auditory biases that likely occur while evaluating the systems and therefore the results of this method may not be valid.

Blind, in-situ methods can have very good bias isolation and are almost natural in terms of spectral, spatial and dynamic attribute judgements and ratings [1,2,4,5,6,7]. But, the basic application of this method is complicated by several practical factors including personal discomfort, requirements for an administrator, and difficulty performing dynamic evaluations.

Dummy head, recordings have also been used to evaluate vehicle sound systems but with limited success [6,7,8]. Binaural recordings made in vehicles and played back over a calibrated system outside the vehicle yield excellent bias isolation, repeatability and spectral representation of the sound system. This method is superior in terms of a fast and easy means to store, retrieve, and compare multiple vehicles (or programs) which is impossible with the other methods. Unfortunately, so much isolation exists in binaural evaluations, that it detrimentally affects several spatial, dynamic, and spectral attributes. This is particularly true for bass reproduction

and, for the reproduction of the proper spatial localization when matching appropriate diffuse inverse filters (for binaural playback) to the listener's personal HRTF characteristics.

As a result, there is a need for a method that offers the freedom and natural aspects of the sighted method, but without the possible bias effects. This paper describes a new sighted in-situ subjective evaluation method for vehicles that is accurate and repeatable, and minimizes the effects of non-auditory biases. It is also reasonable in terms of test time and sound quality attribute resolution. The results of a preliminary study are given.

1.0 SETUP

There were two completely separate phases to this study; phase 1 compared the blind in-situ evaluation method to the placebo method and phase 2 compared the sighted in-situ method to the placebo method. Each phase took approximately 8 weeks to complete and it took nearly 8 months to complete the entire study. Although the variable sets were somewhat different for each phase, the results are still comparable.

1.1 Listeners, Training and Scales

Eight trained listeners with known and acceptable hearing acuity and listening experience were used for this study [5,7,9]. The listeners have had experience evaluating vehicle sound systems using the blind and sighted methods mentioned above. The listeners have successfully completed "ear training and preference testing " exercises [5,6] and were above a 95% accuracy rating with less than 0.5 variance for identifying the correct response.

The listeners were trained to rank and order their judgements using a "fidelity level scale" which is an interval scale ranging from 0-10. Generally, when ranking "high end" vehicle audio systems with this method, there tends to be some scale compression as most of the rankings occur in the 5.0-8.0 range. An example of the scale, with verbal attributes is shown in Table 1. This scale is somewhat similar to the IEC 268-13 interval scale [10].

All listeners were required to "recalibrate" themselves by completing the listening training and fidelity scale testing prior to the vehicle evaluations. An anchor system was available to the listeners throughout the study. The listeners were given explicit instructions about the listening and system calibration procedures.

Rating Range	Verbal Attributes
10	Perfect Reproduction
9	Superior
8	Excellent
7	Very Good
6	Good
5	Fair
4	Poor
3	Bad
2	Terrible
1	Little Resemblance
0	No Resemblance

Table 1. Fidelity Level Scale.

1.2 Vehicle and System Characteristics

Three automobiles equipped with OEM name brand sound systems were used in the phase 1 study and four were used in phase 2. All vehicles were full size sedans with CD/radio combinations. The name brand systems had been evaluated several times in the past and were consistently rated as good-very good sound systems when compared to home systems.

In phase 2, one of the higher end vehicle systems, which was noted for its exceptional sound system, was modified by rolling off the low and high frequencies (to a nearly flat frequency response) to investigate listener bias. All tests were done with the radios set to their detent tone, fade and balance positions. Listening was done in the front driver's position.

1.3 Program Material Characteristics and Selection

A total of ten, 30-60 second stereo programs were chosen for the study with three different types of music, including rock/pop, jazz, and classical. Most of the programs were chosen because they were wide-band, fairly homogenous, and exemplified a wide range of identifiable attributes. Only commercially available program selections were used for this study. A list of the program selections is shown in Table 2.

Recent studies [1,6] have shown that the program material, used for sound quality judgements can have a significant effect on the listener's judgements. Some of the programs were chosen specifically for use in detecting spectral characteristics while others were chosen for their spatial and dynamic characteristics. Most of the program selections used for the evaluations were the same as the listeners used in their training so they were familiar with the content. Peak and linear averaged one-third octave spectrum measurements were performed on all programs to understand the program's spectral characteristics and to allow for comparison with subjective data.

Programs	Phase #	1	2
		CD47a	CD50A
Artist, Title	**Track #**	**EQ#**	**EQ#**
Tracy Chapman, Fast Car	1	0	17
Star Tracks, Battle Star Galactica	2	7	7
Bassissimo, Save it Pretty Mama	3	0	0
Yello, Rubberband Man	4	0	18
R. Ronning, Tieden Bar	5	0	0
Joan Baez, Diamonds and Rust	6	10	0
Little Feat, Hanging on to the Good Times	7	7	0
Delos, Stravinsky, Firebird, Infernal Dance	8	0	0
EBU, Sound Quality Assessment Material, Male speech	9	0	17
Don Randi and Quest, If You Need It	10	10	0

Table 2. Program selections and equalization.

1.4 Automotive Sound System Attributes:

The desired information needed from a listening evaluation can affect the number and type of questions/attributes used and the total time required to complete the evaluation. For instance, in this study, the goal was to compare and rank the performance level of competing audio systems using spectral, spatial and dynamics attributes. Another goal may be to determine specific component and/or system design flaws about a prototype system where several detailed attributes may be required.

One of the functional goals of this study was to maintain a 20-40 minute test time per vehicle for each evaluation round. This becomes important when conducting competitive evaluations between several vehicles (2-4) with several listeners (4-8). Typically, it is desirable to complete the entire evaluation in one day without fatiguing the listeners. A second day may be use to repeat the evaluation to increase the statistical power. The total session time for all four vehicles would then be 80-160 minutes depending on the listeners' abilities. This duration was deemed acceptable for practical evaluation sessions.

The attributes chosen for this study included spectral (quality, balance), spatial (staging, imaging), dynamics (noise, low and high level reproduction) and overall. The overall rating was described to the listeners as a preference rating where they could summarize their general impressions about the system's (audio/environment) performance. A compiled overall attribute rating was derived from the spectral, spatial and dynamic attributes and could be directly compared to the "stated" overall rating to investigate possible bias and other effects.

The program selection's characteristics can substantially contribute to the variance effects in the listeners' judgements and can be more significant than those of the vehicle/system [1,6]. With the intent to understand the variable relations, interactions, and effects on ratings, the subjects were instructed to use the single stimulus method. With this method, each program selection is evaluated for all attributes before proceeding to the next program selection. This method increases the time required to complete an evaluation, but in return, can yield detailed statistics for determining significance and interaction about each program selection. The independent rating method is a key point and imperative to how the placebo method works. Typically, multi-stimulus methods are used for sighted vehicle evaluations where several program selections are evaluated and are "lumped" together for each attribute category.

1.5 Equalization Audibility and Selection

Equalization of the individual program selections was done using a digital editing computer workstation. The workstation, with a DSP farm and other plug-ins, is capable of producing numerous spectral, spatial, dynamic and temporal effects to the program elections with signal processing such as parametric equalization, compression, signal delay, etc. All editing and transfers were maintained in the digital realm. Sixteen equalizations were initially developed in a small listening room with a surround sound system and near field monitors situated to approximate a vehicle's sound field. The spectral equalizations were derived from speaker and automotive sound system characteristics that produced aberrations in the frequency response characteristics such as speaker cone edge holes, comb-filtering effects and poorly aligned filters. A remote controlled parametric equalizer unit with up to 6 filters per channel and parameters ranging from +/- 0.1-15dB amplitude shift, Q's of 0.5 – 10, and center frequency range of 30-15k Hz was used. Signal delays of 3-10ms were also investigated. The equalizations used for this study are shown in Table 3. As many as six filters could cascaded for each equalization.

After the initial equalizations were developed, they were auditioned in several vehicles to determine audibility relative to the program material characteristics. As expected, some the equalizations produced pronounced spectral and spatial effects (time/intensity trading) on some programs while others did not. The equalizations for each program were further modified, to make them more or less audible in the vehicles.

During the equalization phase, the goal was to produce slight to moderate spectral and spatial effects that could be readily detected on a good home sound system in a paired comparison evaluation, but not easily detectable in a vehicle depending on how well the fidelity was maintained. In other words, the modified program equalization effects could be sufficiently masked by the vehicle's environment and/or system transfer function effects. The listener, therefore, had some element of doubt as to the source the aberrations.

Placebo EQ Parameters, Phase 1					Placebo EQ Parameters, Phase 2				
CD#	Freq. Hz.	Amp. dB	1/Q	Delay ms	CD#	Freq. Hz.	Amp. DB	1/Q	Delay ms
7	2k	3	0.3		17				5
	2.7k	-6	0.2		18	650	3	0.2	
	3.8k	4	0.3		7	2k	3	0.3	
	5.5k	-6	0.2			2.7k	-6	0.2	
	8.5k	4	0.4			3.8k	4	0.3	
	11.8k	-6	0.2			5.5k	-6	0.2	
10	100	6	0.5			8.5k	4	0.4	
	200	-8.5	1.1			11.8k	-6	0.2	
	400	6	0.6						
	1k	-4	1						

Table 3. Equalization Parameters.

1.6 Recordings

Because of the desire to maintain a digital only format, CD's were selected as the media of choice. They are widely used in automobiles, offer good bandwidth, dynamic range and ease of use. Several recordable CD's were made for the study with various combinations of equalized and non-equalized selections to study the placebo method. Several CD's were also made with non-equalized programs for comparison to the blind and sighted methods. Each program selection was repeated three times for a total time of approximately 2.5 minutes per selection/track.

Initially, there were several thoughts about how the CD's were to be structured in terms of ratio of equalized to non-equalized programs and total number of programs. Two structures were initially tested: one with repeated programs and equalizations and one with non-repeated programs and repeated equalizations. The repeated structure recordings had four programs and three EQ's each, including flat, (12 programs total) which allowed the listener to hear and compare all of the programs within a single evaluation round. The repeated programs were randomly ordered, but it was found that listeners' memories were good enough to remember the equalized vs. non-equalized differences. Also, with the randomization, an equalized program could be placed either just in front or back of a non-equalized program that would allow easy comparison between the characteristics. For the study, it was decided that the non-repeated program structure would be used to eliminate this problem.

Initially, CD recordings with all ten program selections were used. Four of the ten programs were equalized and six programs were not. Several of the listeners commented that the evaluations were too long and fatiguing. The total time per round took between 40-60 minutes. To shorten the time, the number of programs was reduced to seven, which in turn reduced the round time to about 20-40 minutes.

It should be noted that the equalized to non-equalized ratio doesn't need to exceed 20-30% because the listeners only need to know that some equalized selections are present for the "placebo" effect to work. In fact, as few as 10-20% equalized/non-equalized programs could be used without negatively affecting the results. By itself, the "ego effect" (personal challenge to determine which selections are altered) should reduce the amount of prejudice present in the judgements.

Only the non-equalized data were used to derive the statistics about the listener's ratings on the vehicles. Since the equalized selections were not intended to be used to rate the vehicle's performance, a lower equalized/non-equalized ratio allows for more useful data to rate the vehicles. At this point, it was believed that by comparing the statistics between equalized and non-equalized data, any significant bias effects that occurred between vehicles could be detected.

It should also be noted that this method relies on a two-channel CD source while most vehicles have four channels or quadrants (left front, left back, right front, right back) to deal with. This means that the same equalization effects occurred on either the left-front and left-back channels or right-front and right-back channels simultaneously. This limitation did not seem to constrain the method's usefulness.

1.7 Data Control

Several evaluation forms were developed and tested depending on the number of programs and attributes there were to judge. Both manual (paper forms) and laptop computers with spreadsheet applications were used to collect the data. Mini PC's were found to substantially improve the data collection process. The applications also randomly generated the program selection number for each round. A comment section was also included for each of the attributes where the listener could give more detail about a rating. The listeners were instructed on how to use the software as well as how to best express themselves in the comment section with 3-5 word phrases. The listeners comments are key to understanding the underlying reasons for their ratings. Digital writing pads with templates were also investigated as an input means, but found to be less suitable than typing the information into the laptop PC.

2.0 EVALUATIONS:

The total testing time to complete all evaluation methods was about 8 weeks. The listeners were asked to evaluate 1-2 vehicles per day, and 3 repeats were made on all vehicles and methods for a total of 18 rounds per listener for phase 1 and 24 rounds per listener for phase 2. The repeats were done to increase the statistical power and measure the listener's error variance. Three administrators were used for the blind evaluations and the same administrators were used consistently with the same listeners. The listeners were allowed to repeat the program tracks as often as they needed. The program tracks and vehicle order were randomized for each round.

2.1 Blind In-situ

The blind evaluations were done first, because they provided the most bias isolation. For this method, the vehicles were scented and fitted with seat, steering wheel, and clutch/break covers to reduce the chance of the listener determining what vehicle they were in. The listeners are instructed not to touch any part of the vehicle during the sessions. The listeners started each round in a separate room where they are blindfolded and fit with a pair of headphones (playing

pink noise) to mask any cues from the vehicle or environment. The headphones also allowed communication to the listener while the administrator led them to the vehicle and sat them in the driver's seat.

For the study, all evaluations were made in the driver's seat. Prior to each evaluation the vehicle's volume control was adjusted to an 85dBA level with a pink noise signal recorded on the test CD's. The non-equalized CD's were used for this method. The administrator started the PC software, which returned a random number for the CD program to play. The listener verbally reported the ratings and comments to the administrator, who typed the information into the PC. Each of the seven audio programs was evaluated first for spectral and spatial attributes at the 85dBA level. After these attributes were completed, the administrator repeated the program, and adjusted the volume level as instructed by the listener to test the systems dynamics, low level, and high level attributes. Finally, the listener rated the system for the overall attribute. After the test was completed, the listeners were led back to the room where they started. The average test time per vehicle was 40-70 minutes.

2.2 Placebo

The placebo method was tested after the sighted in-situ method. The listeners were given CD's that contained some equalized programs and were instructed to judge each program as before. The listeners were given explicit instructions that some of the program selections had been modified to alter their spatial and/or spectral characteristics and that they should be listening for changes in the program selection characteristics. Several CD sets, with different labeling, were used and randomized to disguise the program selection characteristics. The listeners entered their ratings and comments on a form or laptop computer. The average test time per vehicle was 20-40 minutes.

2.3 Sighted In-situ

The sighted in-situ evaluations were done using the same program material as used for the blind method. This method was done last as it was the most likely to bias the listeners and allow them to relate the sound characteristics to the vehicles. Time to complete a listening round (one vehicle) ranged from about 20-30 minutes.

3.0 RESULTS: Phase 1-Blind Vs Placebo, Phase 2-Sighted Vs Placebo

The statistical analysis required a null hypothesis, H0, which may be stated as follows: No differences exist between the dependent variable values for the blind, sighted and placebo methods for the attributes tested which cannot be explained by the differences in the independent variables. A significance level of (0.05) was used throughout the analysis. The data from all methods was processed with a statistics program for both parametric and non-parametric methods, including a repeated-measures analysis of variance (ANOVA) as well as the Wilcoxon test [11]. The parametric methods assume a normal Gaussian data distribution, which was generally confirmed for all attributes in both phases of the study. A separate statistical analysis was performed on the data for the non-equalized program selections (noted by the number 2 after the attribute) as well as the combined (equalized and non-equalized) data to observe any significant differences.

Table 4 shows basic descriptive statistics for all attributes used in each method for both phases of the study. The means within each phase are comparable and generally within our target variance of 0.5 points. The standard deviation and error is higher for phase 1 compared to phase 2 and is likely due to adaptation, as the listeners became more familiar with the test methodologies and program selections [7]. It should also be noted that the variance for a vehicle evaluation is generally higher than for home audio equipment evaluations where the comparisons and judgements can be done immediately by switching back and forth between the components [1,4].

Table 4. Descriptive Statistics for each Attribute, Non-equalized Programs

Phase 1, Blind

	Mean	Std. Dev.	Std. Error	Count	Minimum	Maximum	# Missing
Spectral	5.944	1.351	.104	168	2.333	8.667	0
Spatial	5.558	1.576	.122	168	2.000	9.000	0
Dynamic	6.001	1.522	.117	168	3.000	9.000	0
Overall	5.691	1.464	.113	168	3.000	8.500	0
Comp.	5.834	1.350	.104	168	2.778	8.556	0

Phase 1, Placebo

	Mean	Std. Dev.	Std. Error	Count	Minimum	Maximum	# Missing
Spectral.2	5.823	1.297	.100	168	1.333	8.967	0
Spatial.2	5.599	1.363	.105	168	3.000	8.000	0
Dynamic.2	5.678	1.516	.117	168	3.000	8.500	0
Overall.2	5.582	1.247	.096	168	3.000	8.800	0
Comp..2	5.700	1.265	.098	168	2.778	8.111	0

Phase 2, Sighted

	Mean	Std. Dev.	Std. Error	Count	Minimum	Maximum	# Missing
Spectral-Filtered	6.603	.681	.035	384	4.400	8.000	0
Spatial-Filtered	6.458	.843	.043	384	4.100	8.000	0
Dynamics	6.407	.745	.054	192	4.500	8.200	0
Overall	6.510	.664	.048	192	4.700	7.700	0
Comp.	6.517	.563	.041	192	4.689	7.733	0

Phase 2, Placebo

	Mean	Std. Dev.	Std. Error	Count	Minimum	Maximum	# Missing
Spectral-Filtered.2	6.427	.804	.041	384	4.300	8.500	0
Spatial-Filtered.2	6.300	.967	.049	384	4.000	8.600	0
Dynamics.2	6.487	.918	.066	192	4.500	8.500	0
Overall.2	6.423	.798	.058	192	4.700	8.500	0
Comp. .2	6.469	.673	.049	192	4.878	8.222	0

Graphs 1-5 and Tables 5A-5E show the ANOVA results for all main and two-way interactions for the factor Vehicle-by-Method for the Blind and Placebo methods. The results show that there is no significance between the two methods (Blind or Placebo). Both methods demonstrate the same results: Vehicle A is significantly rated the highest while vehicle C is rated second for all of the attributes: Spectral Balance, Spatial Balance, Dynamics, Overall, and Computed Overall.

Graphs 6-10 and Tables 5F-5J the ANOVA results for all main and two-way interactions for the factor Vehicle-by Method for the Sighted and Placebo methods. The results show that there is a definite significance between the two methods (Sighted and Placebo). Both methods showed a distinct lack of preference for Vehicle B (which was re-equalized, so visual bias did not enter into the evaluation). The Placebo method, however, also showed a significant overall preference for Vehicle A, as well as a consistent 2nd, 3rd, and 4th place ranking of the other vehicles. The sighted method could not produce these types of results. This type of "lack of resolution", "blurring", or "confusion" for the sighted method is consistently a characterization of the results for the other attributes. Whereas, for the Placebo method, there is a consistent significance in the data for all the other attributes of Spectral Balance, Spatial Balance, and Dynamics as well.

Graph 11 and Table 6 show the main and interaction results for the spectral attribute in the Blind vs. Placebo study for equalized and non-equalized programs. There are two significant interactions between methods, both of which occur for the equalized programs 2C and 7A. None of the non-equalized programs showed a significant interaction between the methods. It can be seen from this graph that in several cases the equalized program means are greater than the non-

equalized program means, although for the most part they are insignificant due to the high variability. This indicates that the spectral anomalies introduced in the Placebo method were being identified by the listeners. The ANOVA was required to show that they were being identified, which would seem to indicate that the anomalies were subtle, but just noticeable by trained listeners -- not blatant. The method was performing as expected in terms of providing cover of the vehicle's non-auditory attributes, and the listeners were not biasing the results.

Graph 12 and Table 7 show the ANOVA results for the Sighted vs. Placebo Spectral attributes for all programs and vehicles. This information was provided so that the reader could get an idea of the equalized vs. non-equalized program differences. In this case there are several significant interactions taking place for all main effects and the Vehicle by Method interactions. The significance between methods centers primarily around Vehicle C for most Programs. Except for a couple of instances, the sighted means ratings are higher that the placebo ratings. Again, there is a lack of preference for vehicle B for both Sighted and Placebo methods, but the results are somewhat ambiguous between the other Vehicle by Program for both methods. This is likely due to the high variability or lack of statistical power to resolve the differences. It might also indicate that the spectral audibility between the equalized and non-equalized programs was not enough to overcome the system/vehicle effects.

Graph 13 and Table 8 show the ANOVA results for the Sighted vs. Placebo Spatial attributes for all Programs and Vehicles. Again there are several significant interactions taking place for all main and the Vehicle by Method interactions. There is a very significant difference between the Sighted and Placebo methods for Program 1 (spatial shift due to delay) for all Vehicles. There is also significance between methods occurring for most all Programs, again centered around Vehicle C. Vehicle B is again rated the lowest for both methods while vehicle A, C and D are rated nearly the same for the sighted method. Vehicle A and D are a close tie for first with the Placebo method. Again, there is one case (7D) where the equalized program is preferred over the non-equalized program. It is obvious, from this information, that the listeners readily detected a spatial shift, due to the added signal delay. Again, this would indicate that the Placebo method was working as expected.

5.0 CONCLUSIONS

We've seen from the analysis, that the Placebo method is comparable to the Blind method in terms of quality of results. We've also seen that the Placebo method has none of the disadvantages of the Blind method – discomfort, extensive vehicle masking, static only evaluations – and all of the advantages of the Blind method over the Sighted method. The Placebo method is easily managed, can be used with the vehicle in motion, provides statistically significant and repeatable results, and reduces listener bias. We've seen that with well-trained listeners, the Sighted method can provide results that are very similar to those from the Blind or Placebo methods, but there is a lack of resolution and blurring in the sighted results that does not exist for the Blind and Placebo methods. The Placebo method better focuses the listener to the task of listening and in doing so reduces the variability and increases the resolution of the data. With an improved data set, we are provided with a more detailed understanding of the preferences and ratings of our listeners, and we are provided with an increase in the potential for ranking the vehicles beyond first or second ranking levels.

6.0 ACKNOWLEDGMENT

The authors are very appreciative to several people for their work and efforts to make this study possible. This includes, Sean Olive and Floyd Toole, who helped conceive the general concept, Steve Ramsier, who helped with the statistics, our interns; Kevin Updegraff, Eric Boeker, Eric Winninger, and Josh King and all of the listeners.

Table 5. ANOVA tables for non-equalized Blind Vs Placebo and Sighted Vs Placebo methods.

ANOVA Table for Spectral.2 Table 5A

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Vehicle.2	2	4.671	2.336	2.046	.1321	4.093	.405
Method.2	1	.011	.011	.010	.9221	.010	.051
Vehicle.2 * Method.2	2	.366	.183	.160	.8521	.320	.074
Residual	186	212.280	1.141				

ANOVA Table for Spatial.2 Table 5B

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Vehicle.2	2	34.436	17.218	18.312	<.0001	36.624	1.000
Method.2	1	2.210	2.210	2.351	.1269	2.351	.315
Vehicle.2 * Method.2	2	1.020	.510	.543	.5821	1.085	.135
Residual	186	174.891	.940				

ANOVA Table for Dynamic.2 Table 5C

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Vehicle.2	2	26.963	13.481	13.979	<.0001	27.959	1.000
Method.2	1	2.146	2.146	2.226	.1374	2.226	.300
Vehicle.2 * Method.2	2	4.573	2.286	2.371	.0962	4.742	.462
Residual	186	179.373	.964				

ANOVA Table for Comp.2 Table 5D

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Vehicle.2	2	17.680	8.840	12.641	<.0001	25.282	.999
Method.2	1	.002	.002	.003	.9599	.003	.050
Vehicle.2 * Method.2	2	.518	.259	.371	.6909	.741	.107
Residual	186	130.074	.699				

ANOVA Table for Overall.2 Table 5E

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Vehicle.2	2	23.907	11.953	13.274	<.0001	26.547	.999
Method.2	1	.066	.066	.073	.7875	.073	.058
Vehicle.2 * Method.2	2	.064	.032	.035	.9652	.071	.055
Residual	186	167.502	.901				

Tables 5A-5E
Blind Vs Placebo

ANOVA Table for Spectral.2 Table 5G

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Vehicle.2	3	35.010	11.670	22.982	<.0001	68.946	1.000
Method.2	1	5.985	5.985	11.787	.0006	11.787	.948
Vehicle.2 * Method.2	3	4.289	1.430	2.815	.0383	8.446	.675
Residual	760	385.920	.508				

ANOVA Table for Spatial.2 Table 5F

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Vehicle.2	3	200.730	66.910	118.696	<.0001	356.089	1.000
Method.2	1	4.798	4.798	8.511	.0036	8.511	.846
Vehicle.2 * Method.2	3	1.196	.399	.707	.5478	2.122	.196
Residual	760	428.417	.564				

ANOVA Table for Dynamics.2 Table 5H

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Vehicle.2	3	19.420	6.473	9.903	<.0001	29.709	.999
Method.2	1	.618	.618	.945	.3317	.945	.154
Vehicle.2 * Method.2	3	1.655	.552	.844	.4705	2.532	.227
Residual	376	245.783	.654				

ANOVA Table for Overall.2 Table 5I

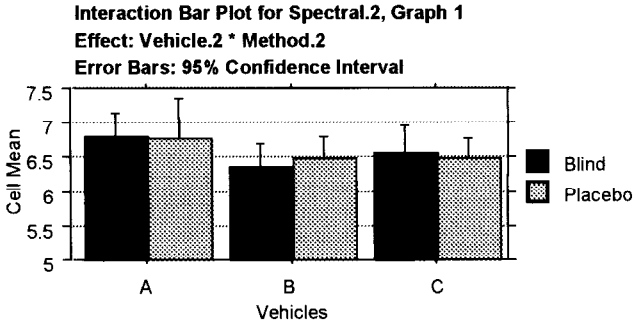
	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Vehicle.2	3	30.791	10.264	22.141	<.0001	66.422	1.000
Method.2	1	.718	.718	1.548	.2142	1.548	.223
Vehicle.2 * Method.2	3	.765	.255	.550	.6485	1.650	.160
Residual	376	174.300	.464				

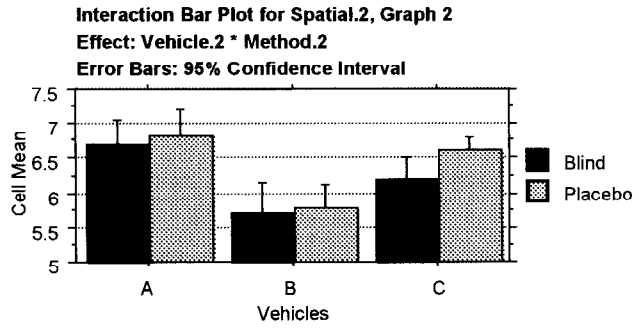
ANOVA Table for Comp. .2 Table 5J

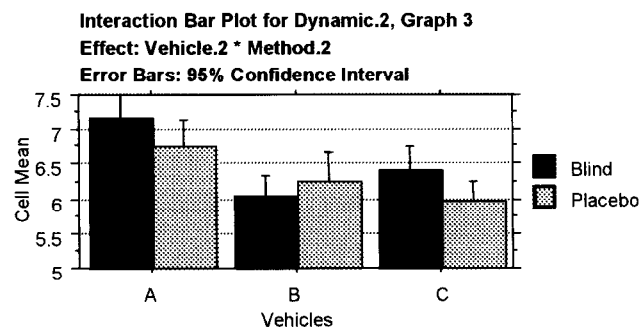
	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Vehicle.2	3	26.474	8.825	27.595	<.0001	82.786	1.000
Method.2	1	.224	.224	.699	.4036	.699	.127
Vehicle.2 * Method.2	3	.451	.150	.470	.7032	1.411	.142
Residual	376	120.241	.320				

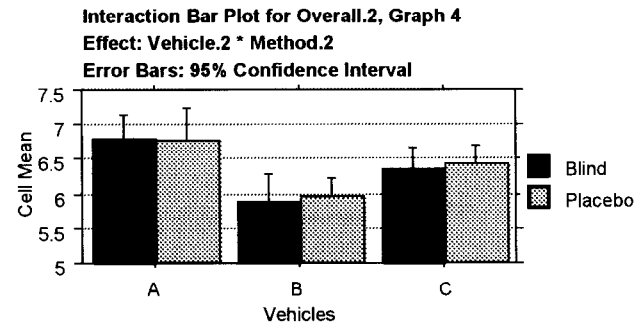
Tables 5F-5J
Sighted Vs Placebo

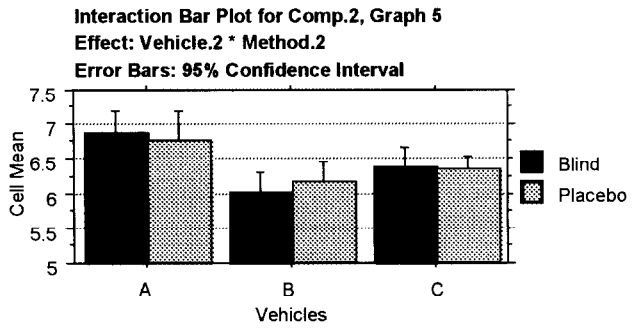
Graphs 1-5. Interaction plots for Blind Vs Placebo, non-equalized programs.



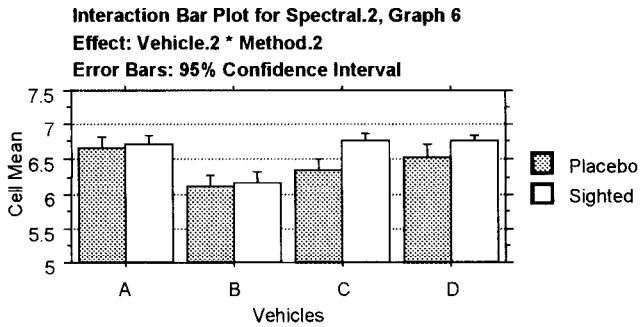


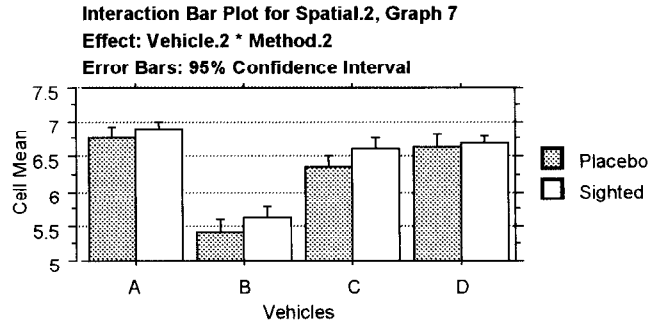


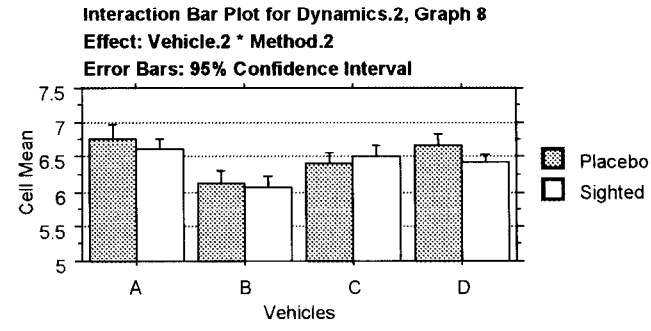


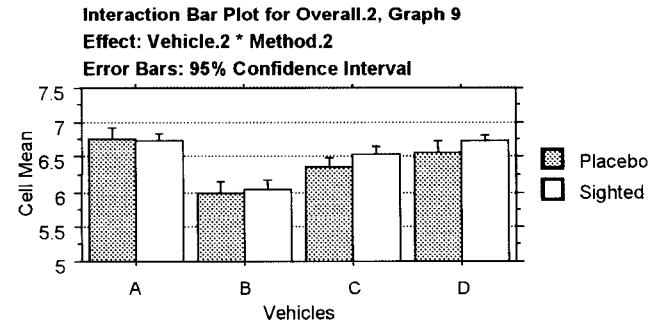


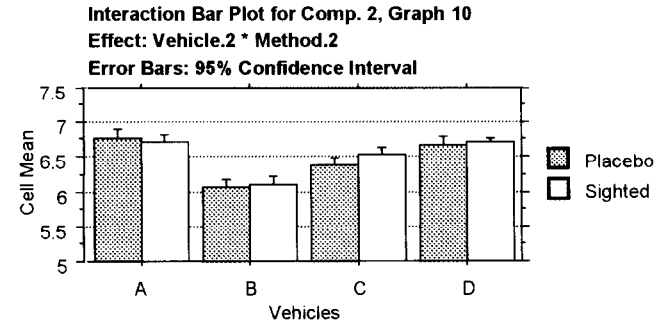
Graphs 6-10. Interaction plots for Sighted Vs Placebo, non-equalized programs.



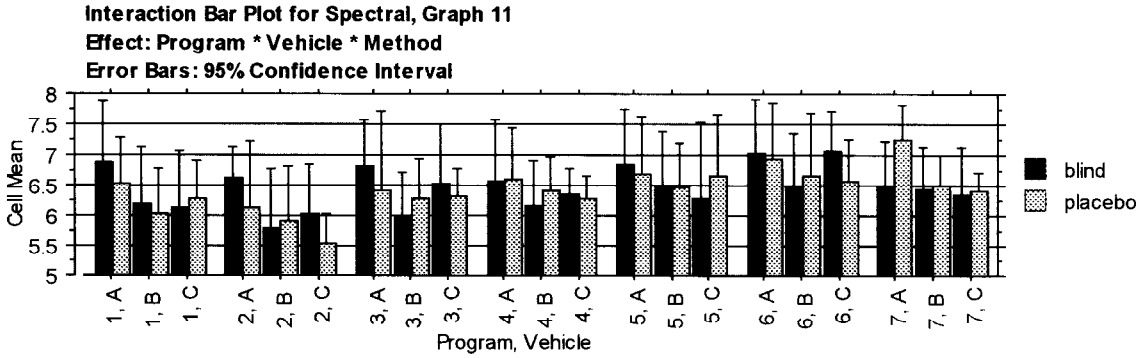








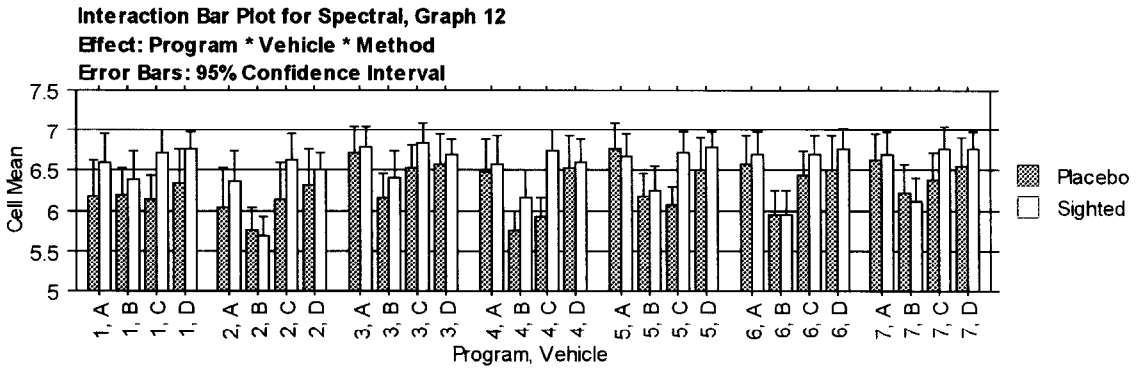
Graph 11, Table 6. Interaction plot for Spectral, Blind Vs Placebo, Programs * Vehicles. The equalized programs are 2, 6 and 7.



ANOVA Table for Spectral, Table 6

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Program	6	20.206	3.368	2.997	.0074	17.982	.910
Vehicle	2	10.913	5.456	4.856	.0084	9.712	.807
Method	1	.299	.299	.266	.6066	.266	.079
Program * Vehicle	12	2.481	.207	.184	.9989	2.208	.118
Program * Method	6	3.193	.532	.474	.8278	2.841	.190
Vehicle * Method	2	.750	.375	.334	.7164	.668	.102
Program * Vehicle * Method	12	5.422	.452	.402	.9623	4.825	.228
Residual	294	330.363	1.124				

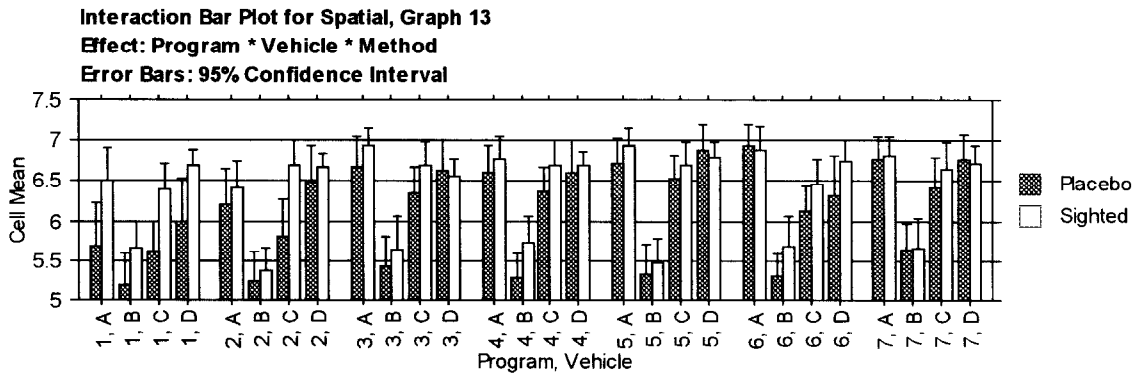
Graphs 12, Table 7. Interaction plot and ANOVA table for Spectral attributes, Sighted Vs Placebo, Programs * Vehicles. The equalized programs are 1,2 and 4.



ANOVA Table for Spectral, Table 7

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Program	6	20.994	3.499	6.010	<.0001	36.058	.999
Vehicle	3	54.999	18.333	31.487	<.0001	94.462	1.000
Method	1	20.110	20.110	34.539	<.0001	34.539	1.000
Program * Vehicle	18	11.130	.618	1.062	.3861	19.116	.766
Program * Method	6	3.074	.512	.880	.5088	5.280	.347
Vehicle * Method	3	8.210	2.737	4.700	.0029	14.100	.908
Program * Vehicle * Method	18	5.937	.330	.566	.9245	10.197	.424
Residual	1288	749.919	.582				

Graphs 13, Table 8. Interaction plot and ANOVA table for Spatial attributes, Sighted Vs Placebo, Programs * Vehicles. The equalized programs are 1,2 and 4.



ANOVA Table for Spatial, Table 8

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Program	6	34.612	5.769	9.037	<.0001	54.225	1.000
Vehicle	3	300.651	100.217	157.007	<.0001	471.020	1.000
Method	1	25.630	25.630	40.154	<.0001	40.154	1.000
Program * Vehicle	18	16.936	.941	1.474	.0904	26.533	.918
Program * Method	6	12.784	2.131	3.338	.0029	20.028	.947
Vehicle * Method	3	3.292	1.097	1.719	.1612	5.157	.441
Program * Vehicle * Method	18	7.284	.405	.634	.8751	11.412	.478
Residual	1288	822.128	.638				

7.0 REFERENCES

[1] F.E.Toole and S.E. Olive, "Hearing is Believing Vs Believing is Hearing: Blind vs. Sighted Listening Test and Other Things," presented at the 97th Convention of the Audio Eng. Soc., San Francisco, (1994 November) Preprint 3894

[2] F.E.Toole, "Subjective Measurements of Loudspeaker Sound Quality and Listener Performance," *J. Audio Eng. Soc.*, Vol. 33, (1985 January/February) pp. 2-32

[3] W.N.House, "Aspects of the Vehicle Listening Environment" Presented at the 87th Convention of the Audio Eng. Soc., New York, (1989 October) Preprint 2873

[4] R.E.Shively, "Subjective Evaluation of Reproduced Sound in Automotive Spaces", in *Proceedings of the AES 15th International Conference* (Copenhagen, Denmark, 1998, Oct 31 – Nov 2), pp 109 –121.

[5] R.E.Shively, W.N.House, "Listener Training and Repeatability for Automobiles," Presented at the 104th Conv. Of the AES, Amsterdam, (1998 May), Preprint 4660

[6] S.E.Olive, "A Method for Training Listeners and Selecting Program Material for Listening Test," Presented at the 97th Convention of the Audio Eng. Soc., San Francisco, (1994 November) Preprint 3893

[7] S.E.Olive, "A Method for Training Listeners: Part II," Presented at the 101st Convention of the Audio Eng. Soc., Los Angeles, (1996 November)

[8] F.E.Toole, "Binaural Record/Reproduction Systems and Their Use in Psychoacoustic Investigations" Presented at the 91st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, Vol. 39, (1991 Dec.), Preprint 3179, p. 1005.

[9] S.Bech, "Selection and Training of Subjects for Listening Test on Sound Reproducing Equipment," *J. Audio Eng. Soc.*, Vol. 40, (1992 July/August) pp.590-610

[10] International Electrotechnical Commission, 268-13: Sound System Equipment; part 13. "IEC Report on Listening Test on Loudspeakers"

[11] SAS Institute Inc., "StatView Reference Manual", 2nd edition, 1998.