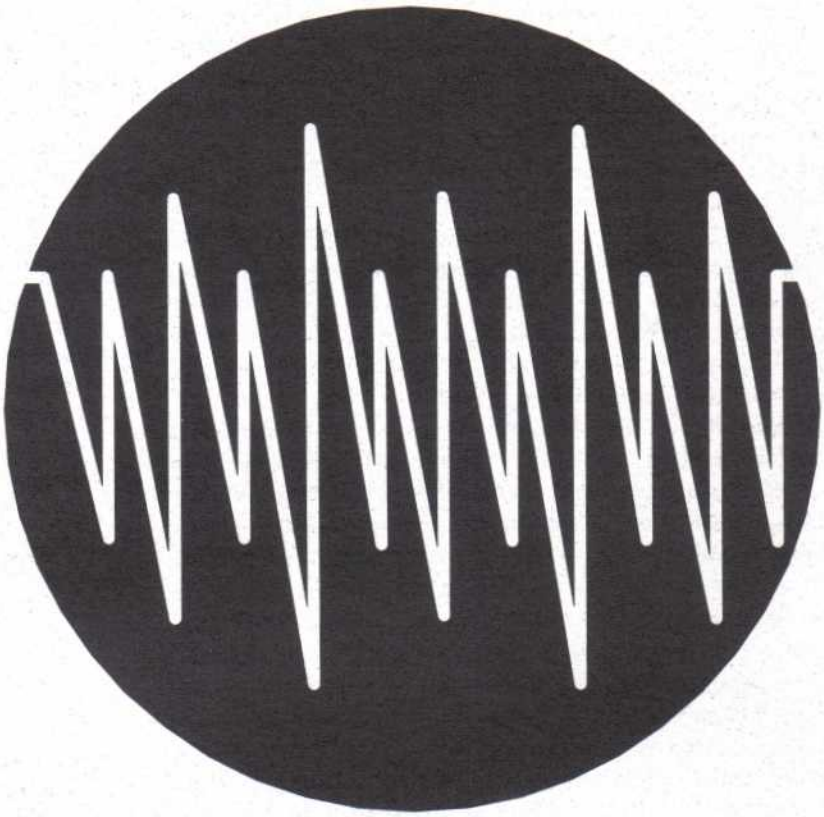
THE SOUND OF MUSIC



Use of a 3-Facet Rasch Model to Measure Small Audio Impairments in the Field of Audio Engineering

David Moulton and *Mark Moulton, Ph.D.*
Moulton Laboratories, Groton, MA, and San Jose, CA

Abstract

The subjective measurement of small audible differences in the audio engineering field has been hampered by experimental conflicts between applicability and reproducibility. The Rasch Model offers a powerful means of controlling the statistical analysis of experimental data in order to maximize reproducibility and applicability across listeners, audio material, and devices under test. The authors describe their testing of five perceptual audio coders for Lucent Technologies.

The Problem of Measuring Perception of Small Audible Impairments

Measurement of listener perception of small audible impairments caused by audio reproduction devices has been constrained by the combined but conflicting needs for (a) reproducible test results and (b) broadly applicable conclusions. Measurement techniques have sought to achieve reproducibility through rigorous test design and execution intended to minimize such sources of uncontrolled variance as listener training and expertise, the choice of program material, and the listening environment. For example, only expert listeners are used and the listening environment must meet exacting specifications. This poses a dilemma. The more rigorously controlled the testing environment, the less faithfully it reflects the listening conditions of the real world. Most listeners are not experts. Most rooms do not meet the specifications for properly controlled listening environments.

Test data are drawn from rating scales such as the Mean Opinion Scale (MOS)[1] and often incorporate an accuracy test in which the listener must pick out a reference signal from among a selection. Analysis of Variance (ANOVA) is used to interpret the results.[2] Data collection rigor is presumed to minimize random statistical variance and to reduce systematic biases. Techniques such as diff-grade analysis are used to diagnose listener inexpertise and to reduce rating scale floor and ceiling effects.[3] However, even under perfectly controlled test conditions, anomalies arise that compromise reproducibility and that ANOVA is not competent to remedy. We encountered several such instances in our study.

The conventional experimental approach is drawn from procedures traditionally used to control objective data from which the human element has been removed. As a consequence it rests on several assumptions that are hard to support. First, it assumes that all extraneous sources of variation can in fact be experimentally removed so that what is revealed are the perceptions themselves and not biases of the listeners, characteristics of the audio systems, or anomalies arising from particular cases. However, the physical and psychological complexity of the listening process appears to render this level of control impossible at the laboratory level. Even under the most controlled conditions, researchers have found replication to be extremely difficult.[4]

Second, it assumes that test subjects unequivocally perceive and can identify the small impairments under test, in other words that they are "experts." Researchers attempt to meet this condition through a process of pre-screening listeners and post-test removal of "non-experts" who fail to meet a guessing accuracy criterion. In reality, of course, listeners bring a continuum of expertise and perceptual acuity to such tests, and no listener is sufficiently expert to produce the kind of reliable measurements ultimately desired. There is also the problem of relating the reports of experts to the probable experience of non-experts. A hypothetical panel of "perfect" experts would lead one to conclude that even the best perceptual audio coding systems are "extremely annoying," leaving fully open the question of how such systems would be perceived by the rest of the world.

Third, there is an assumption that such perceptions can be reduced to a reliable, stable, and reproducible metric, that they are in fact measurable to the point where they may be quantified in a useful way for subsequent use in the design, manufacture, and application of audio systems.[5] It is well known that rating scale data do not possess these metric properties.[6] The relative spacing of the rating scale categories is highly variable and there are pronounced compression effects at the top and bottom of the scale, making it highly nonlinear. While use of diff-grades has made such difficulties more manageable, the fact remains that a rating scale is not a measuring stick.

Fourth, it is assumed that Analysis of Variance is suitable for this type of analysis. However, ANOVA specifies: 1) linear, interval scales; 2) representative samples; and 3) an absence of interaction effects if the intent is to measure main effects. None of these specifications is met in this type of data. The scales are nonlinear. The expert listeners represent no population but their own. Interaction effects abound, and while ANOVA can be used to document their presence, it can do little to prevent their perturbation of the main effects. As a consequence, results drawn from ANOVA do not reproduce well when the selection of programs or listeners is changed.

The Listening Format and Devices Under Test

The devices under test were five high-performance Perceptual Audio Coders known as "codecs." Perceptual Audio Coders are complex encoding algorithms used to remove data from a digital audio signal for ease and speed of electronic transmission. They are "perceptual" in the sense that they take advantage of the physical and psychological mechanics of hearing perception to identify means of removing information from a sound signal in such a way that the brain does not detect the loss. An enormous amount of audio data can be removed before the brain senses anything missing, but eventually as data is removed the brain hears "glitches" in the audio signal. It was the purpose of these tests to measure the audibility of such "glitches" for a specific codec that Lucent Technologies hopes to use in the field of digital radio broadcasting. (Radio broadcasting currently uses "analog" signals which lack the flexibility and wide applicability of digital signals.)

The authors measured the five codecs using a panel of thirty listeners with a wide range of experience (we deliberately included nonexperts) and other demographic characteristics, and ten audio examples drawn from commercial and test recordings. All testing was double-blind and done in small groups over a two-month period, using headphones. The goal of the test was to determine the relative impairment each codec contributed to reference recordings for a range of listeners listening to a range of conventional recordings.

The test consisted of fifty examples, following a training session and three warm-up examples. Each example consisted of a sequence of recordings identified as "Reference", "A," "B," "again, Reference," "A," "B." In each case, the identified reference was one of the Reference recordings, while A or B was the codec-processed copy under test and the remaining of A or B was the reference again (the so-called "hidden reference"). The listeners were asked to score both A and B according to the given criteria, and to identify which of A or B was the hidden reference.

There were two tasks: 1) rating each codec on the 5-point Mean Opinion Scale; 2) picking out the hidden reference. In a conventional diff-grade analysis, the two tasks would be combined into one set of "ratings." The listener would automatically assign a "5" to his *guess* of the hidden reference. The diff-grade would then be the difference between the rating given the *actual* hidden reference and the rating given the encoded signal. These diff-grades would be used to screen out

non-experts. For the Lucent test, listeners were not forced to assign a "5" to one of the choices since diff-grades were not used. Instead, we simply performed two distinct but parallel analyses, the first using the MOS ratings to measure codec transparency, the second using frequency of correct identifications of the hidden reference.

The Mean Opinion Scale was presented as follows:

5 = I cannot hear a difference between the reference and the processed recordings.
4 = I hear a perceptible but not annoying difference between the reference and the processed recordings.
3 = I hear a slightly annoying difference between the reference and the processed recordings.
2 = hear a distinctly annoying difference between the reference and the processed recordings.
1 = I hear an extremely annoying difference between the reference and the processed recordings.

Following the test session, listeners were asked to complete an exit questionnaire. To the question, "Were the PACs in general hard to distinguish from the reference signal?" 27 (90%) answered yes, and 3 (10%) answered no.

Theoretical Justifications for Using a 3-Facet Rasch Model

To analyze the ratings we employed a 3-Facet Rasch Model.[7] Each datum was conceived to be the conjoint effect of the "transparency" of the Codec under test, the "severity" of the Listener, and the "intolerance" of the audio sample or Program to Codec artifacts. The corresponding expression, including an F term to take into account transitions between adjacent categories, was:

$$P\{x_{CLMF} \geq "k" | C_n, L_i, M_j, F_k\} = \frac{e^{C_n - L_i - M_j - F_k}}{1 + e^{C_n - L_i - M_j - F}}$$

where x_{CLMF} = the rating value assigned a Codec
k = a rating scale category
C_n = transparency of Codec n in logits
L_j = severity of Listener i in logits
M_j = intolerance of Program j in logits
F_k = difficulty of the step up from category k-1 to k

Equation 1

In other words, the probability that a given response x will be greater than or equal the k'th rating scale category given Codec C, Listener L, Program M, and step difficulty F of reaching k from k-1, is a function of the logit measures of C, L, M, and F.

The Logit Scale

It will be recalled that conventional subjective testing assumes a stable, linear metric, a condition that is not met by the MOS scale. First, rating scales that have a clear "floor" and "ceiling" such as the MOS scale, whose ratings must fall between "1" and "5," suffer compression effects at the end of the scale. Such effects are ameliorated by using only the middle categories of the scale (not practicable with high-performance codecs) and by using diff-grades, where each rating is replaced by the difference between the rating given the Codec under test and that given a Reference signal. (Diff-grades cleverly smooth out the ceiling effect by introducing the possibility of extra categories at the top of the scale arising from incorrect identifications of the Reference signal, which are then discarded as unreliable, thus locating the set of "reliable" responses towards the center of the diff-grade scale.) The second reason why the MOS metric is not preferred is that, compression effects aside, the length of each rating scale unit depends on the relative wording of adjacent category descriptions, which is highly variable, creating a ruler without consistent units, for which no "centimeter" matches any other.

Rasch measures meet the demand for a stable, linear scale by replacing the MOS rating metric with the logit scale which measures distance in terms of linearized probabilities—the log of the probability of scoring above a specified category divided by the probability of scoring below it. The logit scale suffers no floor or ceiling compression effects as it has no upper or lower limit, and each logit is the same "size" as every other. It can also be readily interpreted as the probability of a particular codec scoring at or above a specified rating when confronted with a listener of a given severity and a program of a given intolerance. Thus, it now appears possible for the audio field to measure perceptual audio coder transparency in a metric as useful and definable as the decibel (which measures loudness on a similarly logarithmic scale) and the other physically defined variables that characterize sound.

Unidimensionality

An important feature of the Rasch Model is that it requires unidimensionality of test items as a condition of fit. Yet all data sets, including the one analyzed here, are multidimensional to some degree, no matter how careful the researchers. What, then, of the Model's applicability? So long as there is a single *dominant* dimension, such as Codec Transparency, the Model is applicable. Extra dimensions manifest as misfit and are purged from the data set accordingly. Thus, unidimensionality is an ideal which the Model tests for and makes it possible to approach. It is not a precondition of successful analysis.

In comparison with the educational and psychological data to which the model is routinely applied, the audio data

set analyzed here was found to be exceptionally unidimensional.

Editing the Data Set to Maximize Reproducibility

Measure reproducibility is the biggest obstacle faced by the audio industry in trying to determine the quality of audio devices. Codecs seem to perform differently in different testing situations no matter how rigorous the testing environment. Part of the problem has been an inability to specify what is meant by reproducibility and to edit data sets to maximize it. The very idea of "editing" a data set sounds heretical from a statistical point of view, and rightly so, not just because ANOVA and other statistical techniques require complete data but because editing compromises the random nature of the sample and thus its representativeness of a larger population. A sample-independent model like Rasch, however, makes no assumptions regarding randomness or representativeness, and it does not require complete data. Indeed, the model is in some senses not a statistical method at all. It merely specifies how data must behave in order to lead to reproducible measures. The data must behave as if attributable to objects that occupy a single position on a single unidimensional scale.

Rasch generates two types of numbers. The first are the logit measures and associated output which correspond to each Codec, Listener, and Program. The second are the expected values expressed in the rating scale metric which are computed for each cell of the data matrix from the logit measures and compared with the corresponding actual data values. It is the summation of their residuals across a set of cells which becomes the basis of the fit statistics associated with each Codec, Listener, and Program.

Suppose, then, we see Codec A misfitting significantly. In conventional Analysis of Variance not much can be done. We can remove Codec A from the analysis, but that leaves us nowhere. We can treat Codec A's overall measure as a Main Effect, then look for the interactions with particular Programs and Listeners that might be causing the misfit and report these as Interaction Effects. But the more pronounced the interaction effects (or biases, for that is what they are), the less trustworthy are the main effects. Could we, then, recompute the main effects after removing the interaction effects? Unfortunately, no. Since ANOVA depends strongly on complete data, on having no missing cells, there is no way to remove the data causing the interaction effects without significantly compromising the interpretability of the results. In short, while ANOVA can offer a diagnosis, it does not supply a cure.

Since it separately models each cell in the data matrix, Rasch does not require complete data. That means the misfitting cells causing interaction effects can be suspended from future analyses (treated as "missing") without compromising the interpretability of the results. The methodology thus implies an iterative process of suspending misfitting data from the analysis (filing it away for diagnostic purposes), recomputing the measures and expected values, identifying and removing the new crop of misfitting cells, recomputing the measures, and so forth. The process is concluded when there are no longer significant misfits, or in other words *when the main effects have been completely purged of interaction effects*.

A full treatment of the relationship of Rasch to ANOVA has yet to be attempted, particularly with respect to the Main Effects/Interaction Effects contrast. I think such would prove enormously valuable to the many fields which, like audio engineering, rely almost exclusively on ANOVA and related methodologies to interpret results. Unable to subtract interaction effects mathematically, researchers must labor to remove them physically from the experiment, often futilely and at great cost.

Results of the Analysis

We performed two parallel and independent analyses, the first to measure codec transparency by a rating scale analysis of the MOS ratings, the second to measure transparency in terms of listener inaccuracy. Since the two forms of analysis are independent methods of looking at the same construct (codec transparency), we felt that a comparison of the two sets of results would act as a cross-check on their reproducibility. Strong agreement would suggest a high likelihood of reproducibility and was in fact found. The correlation between measures derived from the MOS ratings and those derived from the listener's ability to pick the Reference signal in each A/B pair, was 0.99, a pure straight-line relationship, regardless of the fact that the two data sets are substantially independent of each other.

This paper focuses on just the MOS rating scale analysis.

The Rating Scale Analysis

Table 1 gives the MOS generated logit measures for the Codecs and two Reference signals (where the signal suffered no audio coding) after the significant biases were removed (i.e., bias z-score > +2.0 or <-2.0). (The biases themselves and their probable effects on Codec perception will be discussed shortly.) The relative positions of the logit measures in **Table 1** should be very close to those that would be calculated using a different panel of listeners and a different set of audio samples, provided the biases are removed from these as well. The Separation statistic for the codec measures is 8.56, indicating that the codecs have been reliably distinguished by the listening panel. The fact that they are significantly different from the

"Ref" measures (which are computed from ratings given the hidden reference) tells us that the listening panel as a whole was able to reliably detect even the best codecs.

Table 1: Codec measures, biases removed

Codec	Logit Transparency	Model S.E.	Fair Avrge	Misfit
Ref1	3.24	0.17	4.80	1.20
Ref2	3.01	0.15	4.80	1.00
Codec1	2.09	0.11	4.50	1.00
Codec4	2.07	0.13	4.50	1.00
Codec2	1.89	0.12	4.40	1.00
Codec3	1.11	0.10	4.10	0.90
Codec5	-0.28	0.09	3.00	0.90
Mean	1.87	0.12	4.30	1.00
S.D.	1.10	0.03	0.60	0.10

Looking across the top of **Table 1** at the column headings:

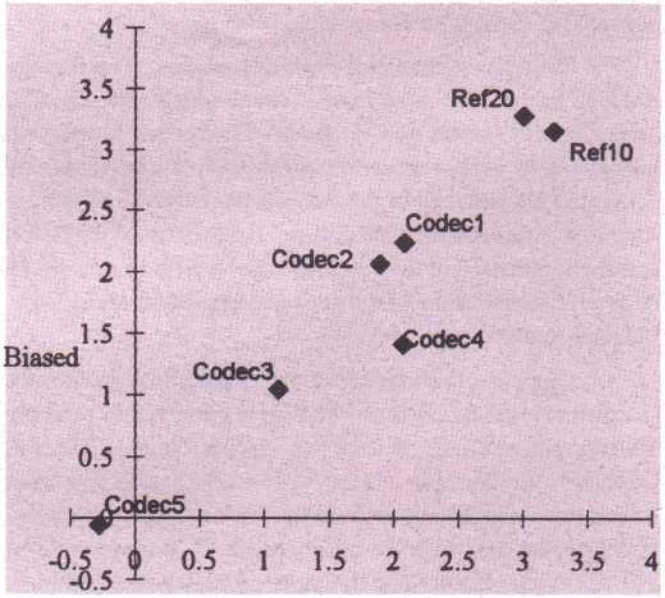
- The **"Codec"** column gives the labels for the codecs analyzed. "Codec4" refers to what we eventually learned was Lucent's PAC at a 96 kb/sec, the audio coder that Lucent plans to use for digital broadcasting. Notice that it performed almost as well as Codec 1 which uses 128 kb/sec, quite a lot more audio information. "Ref1" and "Ref2" are based on the ratings that were given unknowingly to the reference signals when they were compared to Codecs 1 and 2. (Listener comparison of the reference distracters with Codecs 3, 4, and 5 was too easy, artificially inflating their mean MOS scores and creating significant misfit, justifying their exclusion from the analysis.)
- The **"Logit Transparency"** column gives the codec transparency measures on a logit or log-odds unit scale (higher means "more transparent") from which probabilities can be computed using Equation 1. Since the zero point of the scale is arbitrarily set at the mean "severity" level of the Listeners and the mean "intolerance" level of the Programs, the codec probability of scoring "4" (audible but not annoying) or better for the average Listener and Program is easily calculated as: exp(codec measure)/(1 + exp(codec measure)).
- **"Model S.E."** is the standard error in logits of each codec measure, computed assuming the data "fit" the model, an assumption supported by the Misfit column shown next.
- **"Fair Avrge"** is the average of the Rasch expected values for that codec, expressed in the rating scale metric.
- **"Misfit"** is the ratio of observed to expected noise in the estimate and is ideally 1.0. It is calculated as the mean of the squared residuals divided by the variance of the estimate.

On the basis of this table, it was found that Lucent's PAC at 96 kb/sec (Codec 4) was as transparent as Codec 1 which uses 128 kb/sec, a genuine feat of encoding. Applying **Equation 1**, we can say that the average listener (in our sample) listening to the average program will rate Codec 4 "perceptible but not annoying" or better 89% of the time.

Effect of Including Biases

Table 1 is the result of an iterative process of removing biases and interactions between the codecs, programs, and listeners and recalculating parameters. Its virtue is that subsequent analyses with different programs and listeners should result in very similar codec measures, so long as they undergo the same process of removing biases and interactions. However, it does not reveal the peculiarities of this particular test administration. For that, Bias tables (not shown here) are used which show the precise size of the interactions between codecs, programs, and listeners. **Figure 1** graphs the codec measures with biases removed against the codec measures when they have not been removed. You will see that the two sets of estimates are quite similar, with one startling exception.

Figure 1: Biased vs. Unbiased Codec Measures



Unbiased (Rescaled to have same Mean, SD)

Notice that the measure for Codec 4 drops significantly, to third place, when biases are included. A perusal of the bias statistics reveals that approximately 75% of this drop is due to an interaction between Codec 4 and the Castanets audio sample. Reproducing the other samples, Codec 4 performed extremely well. Reproducing an audio sample featuring sharp, percussive castanets, it performed poorly, uncharacteristically so. This provided valuable information to Lucent Technologies, enabling it to identify and remove an error in the encoding algorithm which was causing the castanets interaction.

This raises an important question, of course. Which is the "correct" measure of Codec 4? The answer depends on the goal of the researcher. If the goal is to create reproducible measures, measures which are the same from one testing situation to another, the "correct" measure is the unbiased one—so long as all tests are subject to the same iterative process of removing biases and misfits. If the goal is to describe the effects of a particular testing situation, the biased measure is more reflective of what happened, although it is better to explore biases individually than through their effects on an average.

Listener Severity

The Rasch Model computes estimates for Listeners and Programs at the same time that it estimates Codec transparency. Because the Model makes no assumptions regarding the nature or distribution of the sample, there is no need for Listeners and Programs to be normally distributed along the variable. In fact, it can be seen that the Listener distribution is bimodal, dividing cleanly into "experts" and "non-experts." **Tables 2** and 3 provide Listener and Program measurements.

Table 2: Listener Severity

Listener	Severity	Model	Fair Avrge	Misfit
10/Lro2255	1.04	0.16	2.3	1
5/Gle2255	0.74	0.17	2.6	0.6
8/Mib2253	0.63	0.17	2.7	0.6
1/Eos3255	0.44	0.22	2.8	0.6
2/The3154	0.36	0.18	2.9	0.9
20/Eys3244	0.31	0.18	3	1.2
21/Gla2154	0.25	0.22	3	1.3
22/Nar3144	0.2	0.18	3.1	1.3
24/Har3253	0.2	0.18	3.1	1
28/Dri4243	0.15	0.23	3.1	0.8
30/Moi4255	0.13	0.18	3.1	1
12/Gra1111	0.12	0.19	3.1	1.3
23/Shi2254	0.03	0.19	3.2	1.3
3/Ace4111	0	0.24	3.3	0.8
27/Tin2251	-0.18	0.24	3.4	1.2
25/Urr2113	-0.58	0.22	3.7	0.7
9/Utt2113	-0.63	0.22	3.8	0.6
15/Hra1212	-0.63	0.23	3.8	0.9
26/Gul2133	-0.64	0.23	3.8	0.6
29/Cre2244	-0.65	0.28	3.8	0.9
7/Cou4131	-1.28	0.28	4.2	1.5
Mean	0	0.21	3.2	0.9
S.D.	0.55	0.03	0.5	0.3

Table 2's first column lists each listener with a name abbreviation and background code. The first two digits of the code give their gender (1 = Female) and age (5 = ">50"). The last two digits indicate audio and musical experience where "5" means "extensive training and experience." Notice that the experts cluster toward the top, at the severe end of the scale, the non-experts toward the bottom. Experts are better able to discern audio artifacts, making them more likely to use the lower categories of the scale.

Note also that there are only 21 Listeners listed, though data was gathered for 30. The remaining nine were suspended from the analysis due to high misfit, indicative of internally contradictory response strings. The fact that many of the remaining listeners are non-experts, as evidenced both by their background and their lack of severity, indicates that it is possible to generate reliable measures using non-expert listeners. Because these listener measures are on the same logit scale as the codecs, and because they have been linked to the general population through background demographic information, it becomes possible to make predictions regarding the perception of codecs for the larger population for which they are intended. For instance, taking the average severity measure of those with combined expertise scores of less than 6 as derived from a brief entrance questionnaire, and putting it through **Equation 1**, we find that non-experts (those with little or no musical and audio experience and training) have a 93% chance of finding Codec 4 to be "Perceptible but not annoying" or better. In fact, we can compute the probability that any potential listener will find Codec 4 to be annoying without administering a listening test at all. We need only ask a few questions about musical and audio background and apply a regression equation to predict listener severity, from which probabilities can be computed, a procedure described in another paper.[8]

We can therefore claim that the need for measures having relevance to the larger listening population has been met using only a small, unrepresentative panel of listeners.

Program Intolerance

Finally, let us consider the measurement of Program Intolerance.

Table 3: Program Intolerance to Codec Artifacts

Program	Intolerance	Model	Fair	Misfit
Malespeaking52	0.94	0.14	2.4	1.1
Ethridge1152	0.23	0.13	3	1.1
US3/3342	0.19	0.13	3.1	0.9
B52s2343	0.18	0.14	3.1	0.8
Fagen2233	0.02	0.13	3.2	0.9
Chicago4334	0.01	0.13	3.2	0.9
Sweet Honey2123	-0.08	0.14	3.3	0.9
Castenets4513	-0.15	0.15	3.4	1
Folger2115	-0.52	0.15	3.7	1
Berlioz3115	-0.82	0.16	3.9	0.9
Mean	0	0.14	3.2	0.9
S.D.	0.44	0.01	0.4	0.1

The program column of **Table 3** contains the audio sample used and a code of acoustical characteristics—Dynamic Range, Crest Factor, Distortion, and Reverberence. Interestingly, as the intolerance of the programs to codec artifacts moves up the logit scale, the Reverberence rating decreases from "5" to "1." This suggests that reverberance covers up codec artifacts and can in fact be used to predict program intolerance, just as audio and musical experience can be used to predict listener severity. This was a finding not anticipated by the test administrator. Thus, it is theoretically possible to compute the probability that a given codec will be annoying just by measuring the reverberance of the audio signal electronically.

Observe that the Castanets misfit is a perfect 1.0. This is because its interactions with Codec 4 have been removed. Originally the Castanets misfit was in excess of 1.6.

Conclusions

The Rasch Model shows promise as an inexpensive means of supporting and enforcing the experimental control of audio experiments by statistical means in order to generate reproducible measures. Indeed, in some respects it offers a level of control that extends beyond what could be achieved by ideal experimental conditions, as when it identifies biases and extraneous effects originating from the actual codecs under test. An example of this is the Castanets bias against Codec 4. Since the bias arose from a programming defect within the codec, no amount of experimental control could have prevented it. Without the measurement control imposed by the model, Codec 4's performance would have been doomed to vacillate along the Transparency scale depending solely on the accident of whether or not the Castanets program happened to be present among the sample of programs used in the test.

Are these Rasch measures in fact reproducible? The answer depends on future research, testing the same codecs at a different site using different listeners and programs. There are reasonable grounds for hope. First, we have a well-documented theory supported by extensive educational and psychometric research which finds that such measures will reproduce when a sufficiently diverse set of data have been found to define a coherent variable, i.e., when the data fit the measurement criteria of the model. Second, the reliability statistic for the codec measures, corresponding to a signal to noise ratio of 8.56, is 0.99. Third, a parallel analysis based not on how listeners reported perceiving the codecs, but on their actual success rates in identifying the hidden reference, generates codec measures which are statistically identical ($r = 0.99$) with those generated using the MOS audibility scale, again suggesting reproducibility. We feel that if such preliminary indications are borne out over time, the Rasch Model will prove a useful and cost-saving addition to audio testing methodologies.

Acknowledgments

We wish to acknowledge Søren Beck for his assistance in interpreting the ITU-R testing recommendations during our original research, Benjamin Wright for his advice on research design and use of the Rasch Model, Deepen Sinha for involving us in his codec development work, and Lucent Technologies for their strong ongoing support.

For more information, contact:
Mark Moulton
319-A Page Street
San Jose, CA 95126
(408) 279-1953
E-mail: 73014.340@compuserve.com

1The Mean Opinion Scale is usually a 5-point rating scale (5 = No perceptible difference, 4 = Perceptible but not annoying, 3 = Slightly annoying, 2 = Distinctly annoying, 1 = Extremely annoying). Its use is common in the audio industry.

2ITU-R Recommendation BS.1116. "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," 1996, Section 2. (The ITU is an international board used by the audio engineering field to set standards).

3In the accepted "Reference Signal /Test Signal A /Test Signal B" testing format, one of the two test signals is the reference signal replayed. The listener is supposed to guess which one, then assign it a "5" as having "No audible difference." He then rates the other signal on the 5-point scale. The "diff-grade" is the difference between the rating assigned the hidden reference and that assigned to the other test signal (Diff-grade = Hidden Reference rating - Other Test Signal rating). This is the metric recommended by the ITU. For more on diff-grades, see Section 3.0.

4Thomas Sporer, "Evaluating Small Impairments with the Mean Opinion Scale—Reliable or Just a Guess?" AES Preprint, November 1996.

5ITU-R Reccomendation BS.1116, Section 3.2.

6Benjamin Wright & Geoffrey Masters, *Rating Scale Analysis* (Chicago: MESA Press, 1982).

7 John M. Linacre, *Many-Facet Rasch Measurement* (Chicago: MESA Press, 1994), pp. 1-21.

8We describe just such an analysis performed using these data. David and Mark Moulton, "Codec 'Transparency,' Listener 'Severity,' Program 'Intolerance': Suggestive Relationships between Rasch Measures and Some Background Variables." Audio Engineering Society Preprint, 106th AES Convention, September 1998, San Francisco.

David Moulton:
Audio Engineer, Educator, and Author. Owner of Digital Media Services, a multitrack and surround post-production facility. Principal in Sausalito Audio Works, a firm licensing high-performance loudspeaker technology. Author of "Golden Ears Audio Ear Training" and "Total Recording" (1998 release). Degrees from Bard College and Juilliard School of Music.

Mark Moulton, Ph.D.:
Psychometrician. Specializes in statistical measurement and prediction using Rasch analytic techniques in a variety of fields including psychology, audio, and economic forecasting. Author of "n-Dimensional Replacement: Implications of a Rasch Geometry." Degrees from St. John's College and University of Chicago.