International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

Report ITU-R BS.2300-0

(04/2014)

Methods for Assessor Screening

BS Series

Broadcasting service (sound)

International Telecommunication Union

Foreword

The role of the Radiocommunication Sector is to ensure the rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including satellite services, and carry out studies without limit of frequency range on the basis of which Recommendations are adopted.

The regulatory and policy functions of the Radiocommunication Sector are performed by World and Regional Radiocommunication Conferences and Radiocommunication Assemblies supported by Study Groups.

Policy on Intellectual Property Right (IPR)

ITU-R policy on IPR is described in the Common Patent Policy for ITU-T/ITU-R/ISO/IEC referenced in Annex 1 of Resolution ITU-R 1. Forms to be used for the submission of patent statements and licensing declarations by patent holders are available from http://www.itu.int/ITU-R/go/patents/en where the Guidelines for Implementation of the Common Patent Policy for ITU-T/ITU-R/ISO/IEC and the ITU-R patent information database can also be found.

Series of ITU-R Reports

(Also available online at http://www.itu.int/publ/R-REP/en)

Series	Title
BO	Satellite delivery
BR	Recording for production, archival and play-out; film for television
BS	**Broadcasting service (sound)**
BT	Broadcasting service (television)
F	Fixed service
M	Mobile, radiodetermination, amateur and related satellite services
P	Radiowave propagation
RA	Radio astronomy
RS	Remote sensing systems
S	Fixed-satellite service
SA	Space applications and meteorology
SF	Frequency sharing and coordination between fixed-satellite and fixed service systems
SM	Spectrum management

***Note**: This ITU-R Report was approved in English by the Study Group under the procedure detailed in Resolution ITU-R 1.*

Electronic Publication
Geneva, 2014

REPORT ITU-R BS.2300-0

Methods for Assessor Screening

(2014)

Summary

This Report contains a description of methods for the screening of experienced assessors in Report ITU-R BS.1534 and related listening tests. The expertise gauge (eGauge) method describes in detail a means of rapidly and robustly selecting experienced assessors. Software for this method is available on:

ITU-R eGauge 7.3.zip

TABLE OF CONTENTS

Page

1 Introduction

Report ITU-R BS.1534 advises that experienced assessors be used in order to collect high quality listening test data. This Report describes methods for the selection of experienced assessors. The "expertise gauge" (eGauge) method [1] describes in detail a means of rapidly and robustly selecting experienced assessors. Software for the method is available on:

ITU-R eGauge
7.3.zip

This Report focuses upon methods for the screening of experienced assessors for usage with Report ITU-R BS.1534 and related recommendations. The method seeks to efficiently identify experienced assessors that are suitable for inclusion in data analysis based upon the following assumptions:

– assessor experience is to be shown within an experiment (a pilot study or the main experiment);

– data from Report ITU-R BS.1534 experiments are to be treated as absolute in nature;

– assessor experience is to be demonstrable based on a minimum of one attribute.

An experienced assessor is chosen for his/her ability to carry out a listening test. This ability is to be qualified in terms of the assessors Reliability and Discrimination skills within a test, based upon replicated evaluations.

The expertise gauge (eGauge) approach measures three performance characteristics, in relation to assessor ratings as illustrated in Fig. 1.

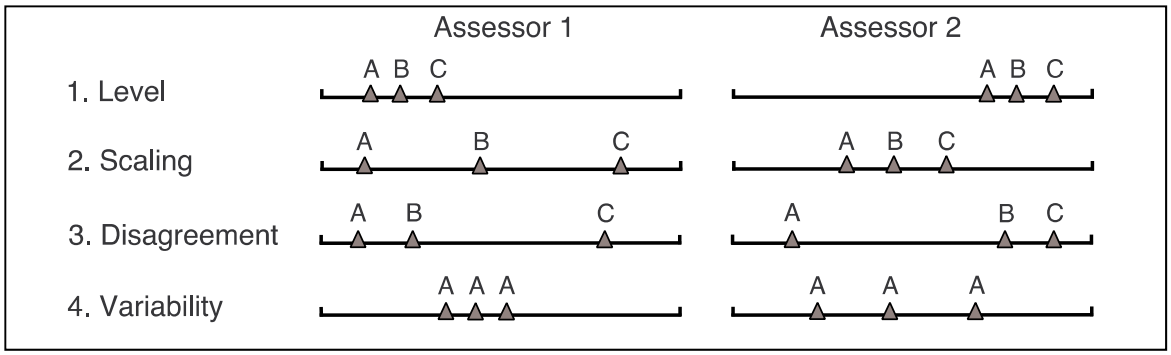
– Discrimination: a measure of the ability to perceive differences between test items.

– Reliability: a measure of the closeness of repeated ratings of the same test item.

– Panel Agreement: a measure of the closeness of ratings between a listener and the panel.

FIGURE 1

The four basic assessor differences in scale ratings. Letters A, B and C represent the scores of three different systems



The method considers the overall performance of the assessor in the evaluation of all test stimuli (systems and samples), excluding anchors of reference samples.

The three test metrics of discrimination, reliability and agreement are calculated based upon an analysis of variance of the data. A non-parametric permutation test is then applied to each metric to define a threshold of acceptability and provide a robust method for the performance categorization of assessors within any given test. Based upon the analysis of discrimination and reliability performance for test stimuli, it is possible to objectively quantify and establish what category an assessor's performance falls into, in accordance with ISO 8586-2 [3] (see Table 1).

For the needs of Report ITU-R BS.1534, assessors with performance falling below the permutation test level for both discrimination and reliability will be categorized as *naïve*, and as such can be excluded from the test analysis. Assessor exceeding the permutation test level for both discrimination and reliability may be categorized as *selected or experienced assessors*.

TABLE 1

Assessor categorization terminology based upon ISO 8586-2 [3]

Assessor category	Performance description
Assessor	Any person taking part in a sensory test
Naïve assessor	A person who does not meet any particular criterion
Initiated assessor	A person who has already participated in a sensory test
Experience assessor (selected assessor [3])	Assessor chosen for his/her ability to carry out a sensory test
Expert assessor	Selected assessor with a high degree of sensory sensitivity and experience in sensory methodology, who is able to make consistent and repeatable sensory assessments of various products

2 Technical descriptions

The model described herein is an evolution of the original expertise Gauge (eGauge) approach developed, tested and reported in [1].

The eGauge model proposed here has been improved in a number of ways. Primarily, the new model is able to handle both 4- or 5-factor datasets as commonly encountered in Report ITU-R BS.1534 tests. Typically 4-factor experiments comprise systems, samples, replicates and assessors. 5-factor experiments may have an additional factor, generically referred to as condition. “Condition” may refer to important experimental characteristics such as bitrate or other parameters.

The method uses an ANOVA (analysis of variance) of the 4- or 5-factor data to calculate the three performance metrics, namely, discrimination, reliability and agreement.

An unfolding methodology is applied on the data in order to reduce the number of factors in the ANOVA model. From a 2-way (system, sample) or 3-way (system, sample, condition) ANOVA, the factor/column system, sample and condition are merged to create a new factor: stimuli. The factor stimuli is equivalent to:

System + Sample + (Condition) + System * Sample + (System * Condition + Sample * Condition + System * Sample * Condition).

Therefore the explained variance of *stimuli* is actually the variance explained by the experimental design.

In the following description the variables are:

– *k* is a replicate between 1 and *K*;

– *i* is a stimuli between 1 and *I*;

– *j* is an assessor between 1 and *J*.

After the unfolding, the following values are extracted:

– count *K*, the number of replicates;

– calculate *Xi* the average value of each stimulus.

The following calculation is run on each assessor:

– compute a 1-way ANOVA in order to get the mean square error (MSE*j*) and the mean square from the stimuli factor (MSS*j*);
– calculate *Xij* the average value of each stimulus;
– calculate the SPAN*j*, the average standard deviation of a score given by the assessor *j*;
– calculation of the sum of square of the Disagreement MSD*j*.

From these values, the reliability, discrimination and agreement are computed:

– reliability *j* is the SPAN (average of all the SPAN*j*) divided by the mean square error of assessor *j* from the ANOVA model;
– discrimination *j* is a *F*-value, it is the ratio between the MSS*j* and the MSE*j*;
– agreement is the ratio between the SPAN and the MSD*j*.

The three metrics, reliability, discrimination and agreement provide an overview of the assessor performance. A non-parametric permutation test [4] is then used as a test of significance. The permutation test is computed using 150 iterations per assessor, in which the systems are shuffled per assessor in each replicate for the calculation of the reliability and discrimination. This is repeated for all assessors to calculate the permutation test level of the test.

For agreement, the data of one assessor are shuffled one at a time and compared to the overall panel and this operation is iterated for each assessor to calculate the permutation test level of the test.

In practical terms the permutation test defines the so-called noise floor of the assessor performance for reliability and discrimination metrics. Below this level, assessor performance is equivalent to random ratings, which only degrade the quality of the data and the estimates of central tendency.

3 Example output and assessor screening

The eGauge method provides four graphs as output. The three metrics (discrimination, reliability and agreement) are plotted as bar graphs for each assessor (Fig. 5). The black line in each plot indicates the non-parametric permutation test level. Additionally, a summary scatter plot is provided of reliability versus discrimination (see Fig. 6). This Figure has four quadrants delineated by the permutation test levels for the two eGauge metrics: reliability and discrimination. The quadrants are illustrated in Fig. 2 and explained in Table 2.

TABLE 2

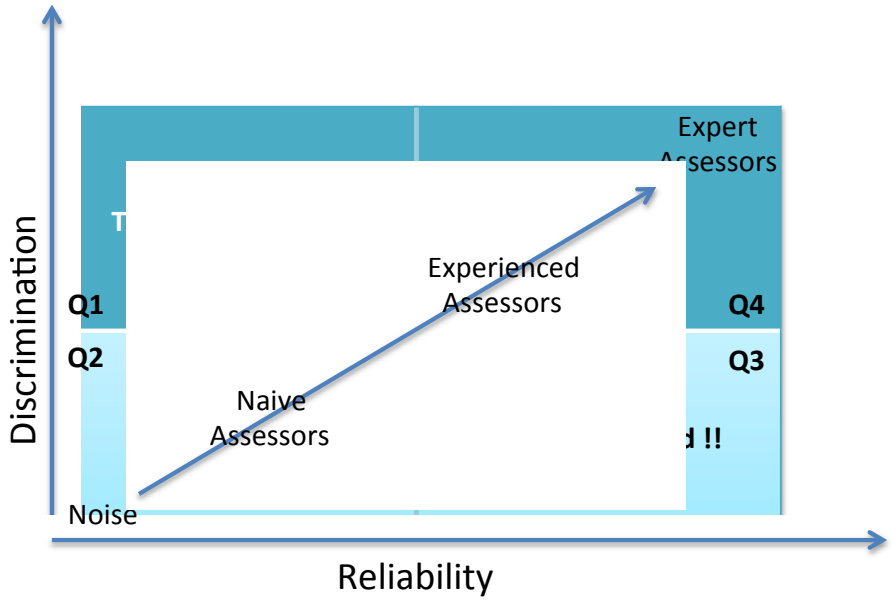
Description of quadrant definitions and actions for eGauge reliability and discrimination scatter plots

Quadrant	Assessor performance description	Categorization	Action
Quadrant 1	Good discrimination, Poor reliability skills	Naïve assessor	Training required Exclude from analysis
Quadrant 2	Poor discrimination, Poor reliability skills	Naïve assessor	Training required Exclude from analysis
Quadrant 3	Poor discrimination, Good reliability skills	Naïve assessor	Training required Exclude from analysis
Quadrant 4	Good discrimination, Good reliability skills	Experienced (or selected) assessor	Include in analysis

Assessors in the top right of quadrant 4 show a high degree of expertise in Fig. 2.

FIGURE 2

Quadrant description for eGauge scatter plot of reliability versus discrimination. The permutation test level for the two metrics provides the delineation between quadrants



The agreement plot is informative regarding the degree of agreement between assessors. Assessors below the permutation test level are in poorer agreement with the panel mean compared to assessors above the permutation test level.

Once the data has been analysed, it is possible to select and report suitably experienced assessors for inclusion in the final analysis. Assessors whose discrimination and reliability ratings exceed the permutation test level (defined by the dark line in Figs 3 and 4) shall be considered as *experienced assessors* for the purposes of the experiment under analysis. Assessors are categorized as naïve if their rating on either or both reliability or discrimination metrics fall below the permutation test threshold and will be excluded from the analysis.

FIGURE 3

eGauge assessor discrimination plot

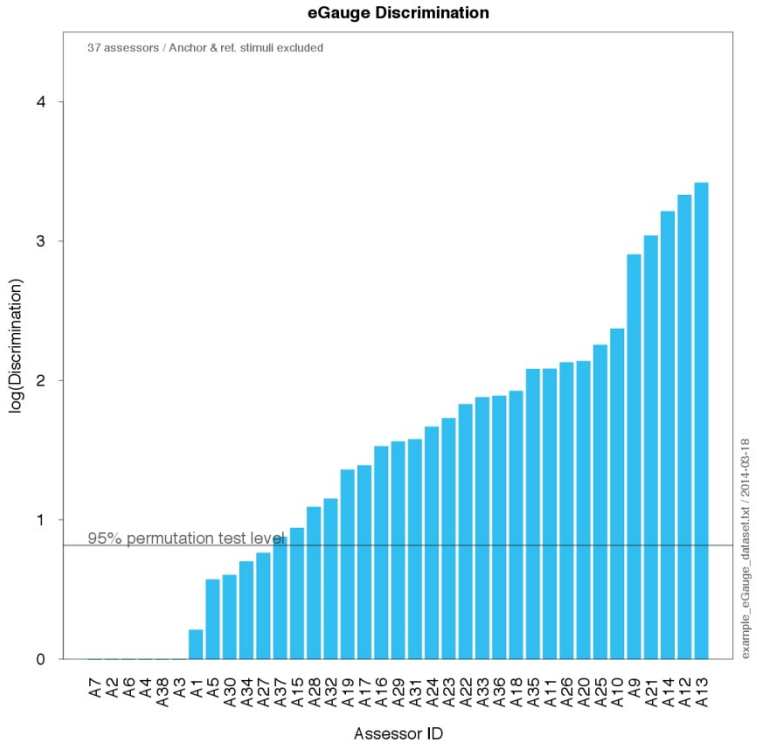


FIGURE 4

eGauge assessor reliability plot

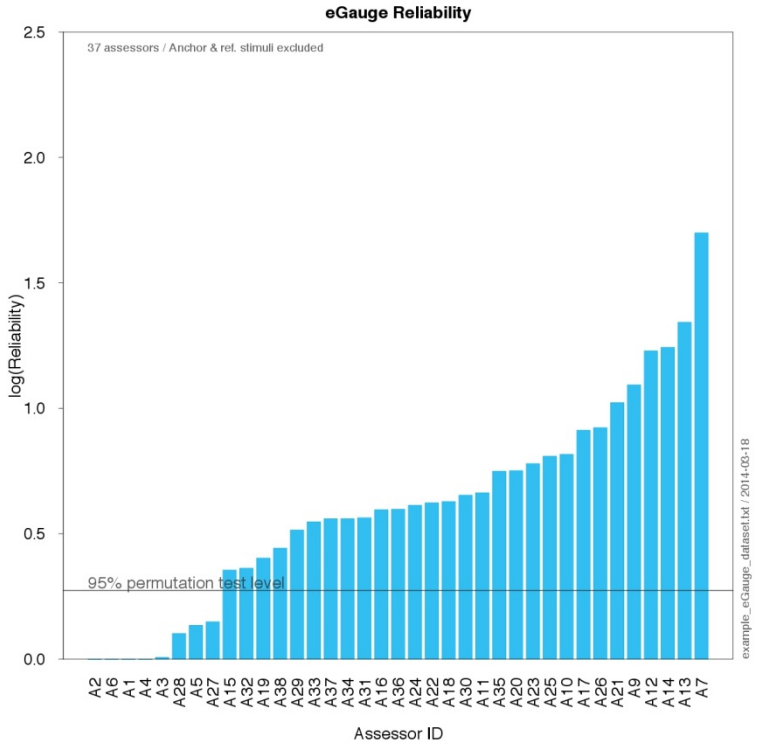


FIGURE 5

eGauge panel agreement plot

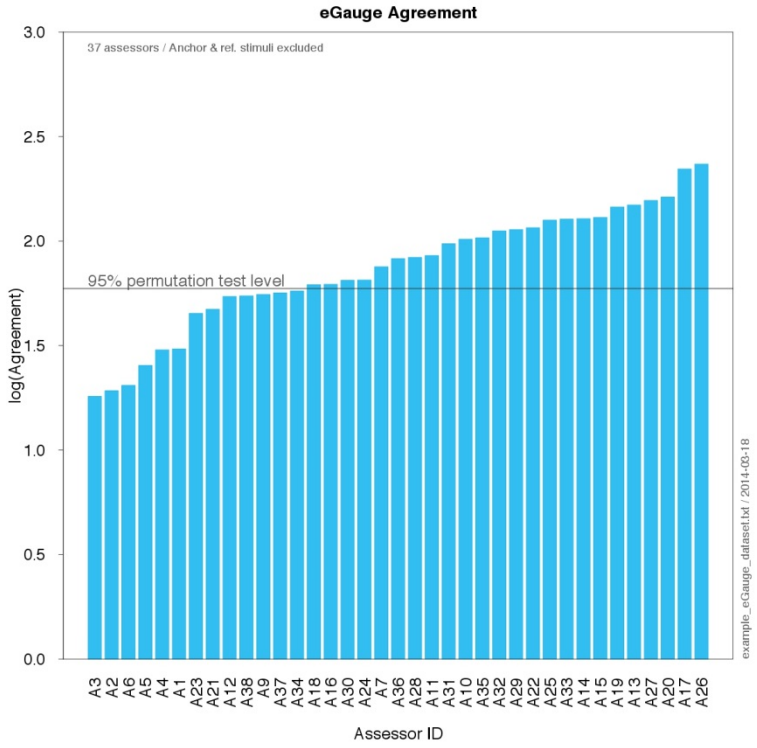
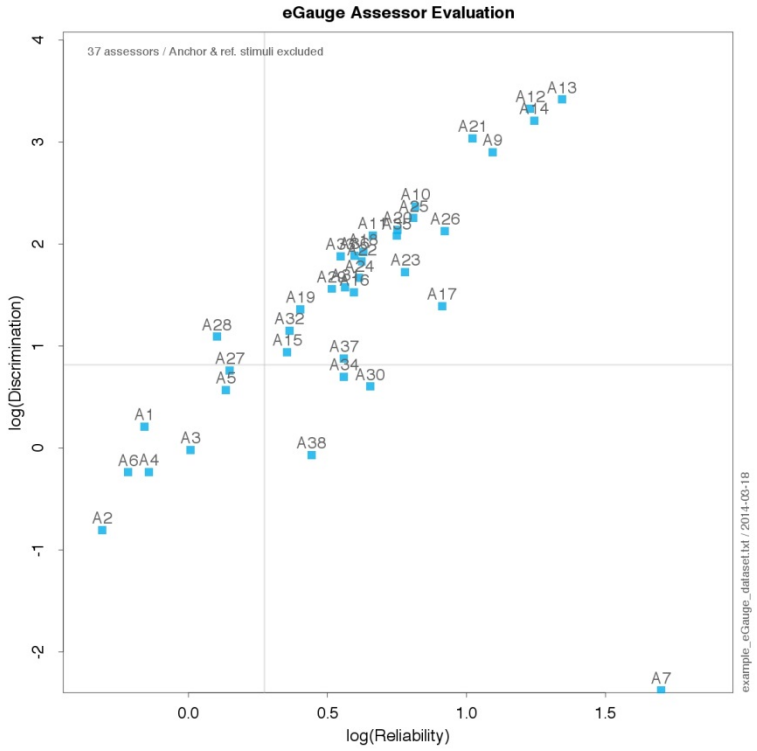


FIGURE 6

Combined eGauge assessor reliability and discrimination plot



4 Results for inclusion in test report

All four output plots may be provided in the test report to demonstrate the degree of assessor experience. Only data from qualified experienced assessors in pre- or post-screening should be included in test data analysis. Assessors should be anonymised in the test report.

If pre-screening pilot experiment was performed, a full description of this pilot study should be provided to demonstrate its validity of the stimuli for the screening and categorization of assessors for the main experiment.

5 Source code

The stable source R (for R version 3.0.1) code for eGauge is available on:

ITU-R eGauge 7.3.zip

The open source R environment for statistical analysis is available from: http://cran.r-project.org

6 Common listening tests data format

The data structure proposed here should be sufficiently generic to allow for analysis of data from Report ITU-R BS.1534 test data. Additionally, the format allows for import to all commonly employed statistical analysis tools and environments, such as SPSS, SAS, Matlab, XLStat, R, etc.

Data shall be stored in a tab delimited text file (.txt) and will employ a "." as the decimal separator. This format can be directly imported into Microsoft Excel as well and other common statistical analysis tools for editing and manipulation.

Each row should be the evaluation of one stimulus by one assessor for one replicate.

The first row of the file shall contain the column labels for all the data, according to the following definitions:

TABLE 3

Common listening tests data format structure

Header	**AssessorID**	**SystemID**	**SystemLabel**	**SampleID**	**SampleLabel**	**ConditionID**	**Condition Label**	**Replicate**	**Rating**
Description	Assessor identification	System number	Test system name	Sound sample number	Sound sample name	Optional additional test factor number	Optional additional test factor name (e.g. bitrate)	The replicate number	Assessor rating
Type	Text string	Numeric	Text string	Numeric	Text string	Numeric	Text string	Numeric	Numeric
Details		Reference = 0 Anchor = –1, –2, etc.		Use 1 to *n*		Use 1 to *n*		Use 1 to *n*	Use "." as decimal separator

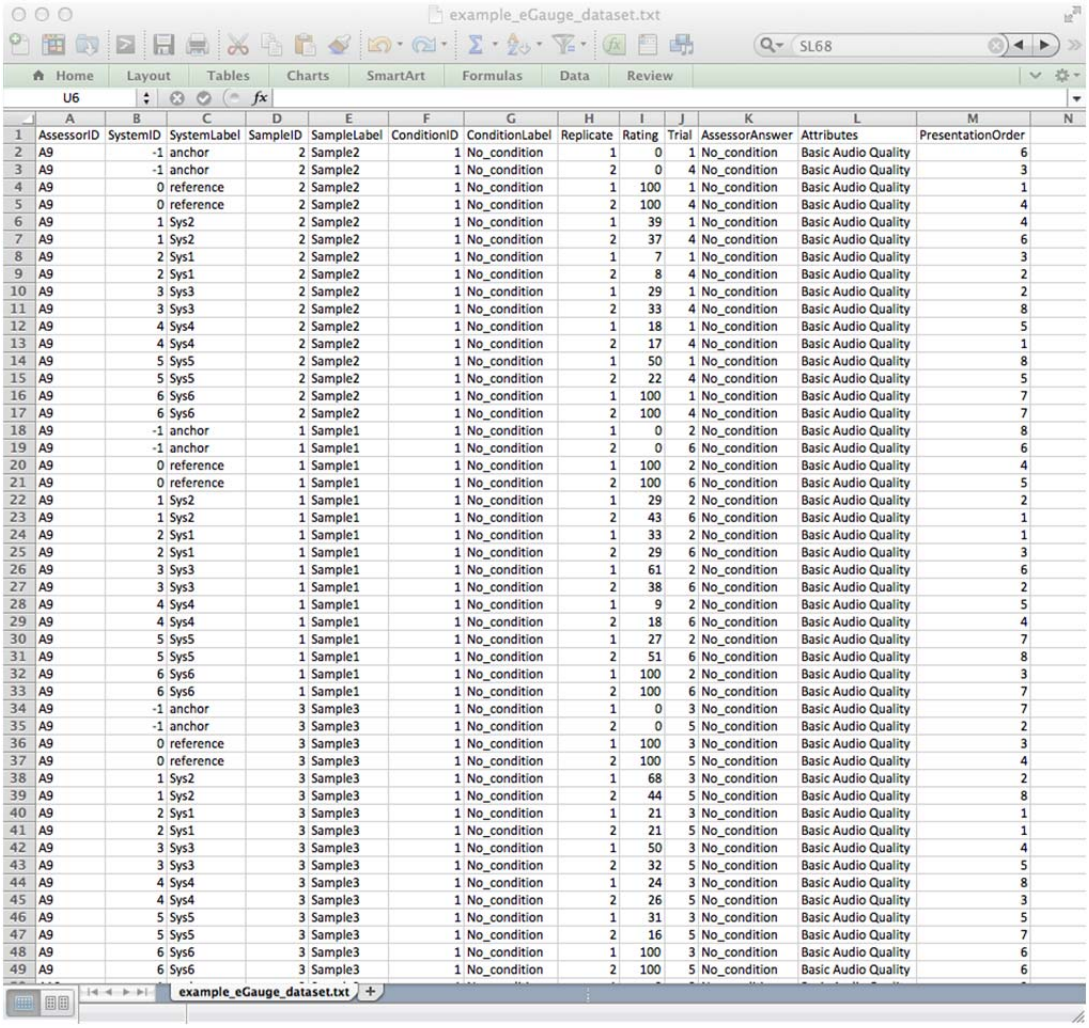
Column header labels are case sensitive.

The SystemID of the reference should be 0 and the SystemID of the anchor should be –1. In the case of additional anchors, these will be labelled with a negative SystemID, e.g. –2, –3, etc.

If one or more factors are not used in the experiments they should however be in the data. The numeric ID and the label should then have only one level. See the factor "condition" in the following example (see Fig. 7).

6.1 Example data format

FIGURE 7
Example common listening tests data format, when imported into Microsoft Excel (.xls).



example_eGauge_dataset.txt

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	AssessorID	SystemID	SystemLabel	SampleID	SampleLabel	ConditionID	ConditionLabel	Replicate	Rating	Trial	AssessorAnswer	Attributes	PresentationOrder	
2	A9	-1	anchor	2	Sample2	1	No_condition	1	0	1	No_condition	Basic Audio Quality	6	
3	A9	-1	anchor	2	Sample2	1	No_condition	2	0	4	No_condition	Basic Audio Quality	3	
4	A9	0	reference	2	Sample2	1	No_condition	1	100	1	No_condition	Basic Audio Quality	1	
5	A9	0	reference	2	Sample2	1	No_condition	2	100	4	No_condition	Basic Audio Quality	4	
6	A9	1	Sys2	2	Sample2	1	No_condition	1	39	1	No_condition	Basic Audio Quality	4	
7	A9	1	Sys2	2	Sample2	1	No_condition	2	37	4	No_condition	Basic Audio Quality	6	
8	A9	2	Sys1	2	Sample2	1	No_condition	1	7	1	No_condition	Basic Audio Quality	3	
9	A9	2	Sys1	2	Sample2	1	No_condition	2	8	4	No_condition	Basic Audio Quality	2	
10	A9	3	Sys3	2	Sample2	1	No_condition	1	29	1	No_condition	Basic Audio Quality	2	
11	A9	3	Sys3	2	Sample2	1	No_condition	2	33	4	No_condition	Basic Audio Quality	8	
12	A9	4	Sys4	2	Sample2	1	No_condition	1	18	1	No_condition	Basic Audio Quality	5	
13	A9	4	Sys4	2	Sample2	1	No_condition	2	17	4	No_condition	Basic Audio Quality	1	
14	A9	5	Sys5	2	Sample2	1	No_condition	1	50	1	No_condition	Basic Audio Quality	8	
15	A9	5	Sys5	2	Sample2	1	No_condition	2	22	4	No_condition	Basic Audio Quality	5	
16	A9	6	Sys6	2	Sample2	1	No_condition	1	100	1	No_condition	Basic Audio Quality	7	
17	A9	6	Sys6	2	Sample2	1	No_condition	2	100	4	No_condition	Basic Audio Quality	7	
18	A9	-1	anchor	1	Sample1	1	No_condition	1	0	2	No_condition	Basic Audio Quality	8	
19	A9	-1	anchor	1	Sample1	1	No_condition	2	0	6	No_condition	Basic Audio Quality	6	
20	A9	0	reference	1	Sample1	1	No_condition	1	100	2	No_condition	Basic Audio Quality	4	
21	A9	0	reference	1	Sample1	1	No_condition	2	100	6	No_condition	Basic Audio Quality	5	
22	A9	1	Sys2	1	Sample1	1	No_condition	1	29	2	No_condition	Basic Audio Quality	2	
23	A9	1	Sys2	1	Sample1	1	No_condition	2	43	6	No_condition	Basic Audio Quality	1	
24	A9	2	Sys1	1	Sample1	1	No_condition	1	33	2	No_condition	Basic Audio Quality	1	
25	A9	2	Sys1	1	Sample1	1	No_condition	2	29	6	No_condition	Basic Audio Quality	3	
26	A9	3	Sys3	1	Sample1	1	No_condition	1	61	2	No_condition	Basic Audio Quality	6	
27	A9	3	Sys3	1	Sample1	1	No_condition	2	38	6	No_condition	Basic Audio Quality	2	
28	A9	4	Sys4	1	Sample1	1	No_condition	1	9	2	No_condition	Basic Audio Quality	5	
29	A9	4	Sys4	1	Sample1	1	No_condition	2	18	6	No_condition	Basic Audio Quality	4	
30	A9	5	Sys5	1	Sample1	1	No_condition	1	27	2	No_condition	Basic Audio Quality	7	
31	A9	5	Sys5	1	Sample1	1	No_condition	2	51	6	No_condition	Basic Audio Quality	8	
32	A9	6	Sys6	1	Sample1	1	No_condition	1	100	2	No_condition	Basic Audio Quality	3	
33	A9	6	Sys6	1	Sample1	1	No_condition	2	100	6	No_condition	Basic Audio Quality	7	
34	A9	-1	anchor	3	Sample3	1	No_condition	1	0	3	No_condition	Basic Audio Quality	7	
35	A9	-1	anchor	3	Sample3	1	No_condition	2	0	5	No_condition	Basic Audio Quality	2	
36	A9	0	reference	3	Sample3	1	No_condition	1	100	3	No_condition	Basic Audio Quality	3	
37	A9	0	reference	3	Sample3	1	No_condition	2	100	5	No_condition	Basic Audio Quality	4	
38	A9	1	Sys2	3	Sample3	1	No_condition	1	68	3	No_condition	Basic Audio Quality	2	
39	A9	1	Sys2	3	Sample3	1	No_condition	2	44	5	No_condition	Basic Audio Quality	8	
40	A9	2	Sys1	3	Sample3	1	No_condition	1	21	3	No_condition	Basic Audio Quality	1	
41	A9	2	Sys1	3	Sample3	1	No_condition	2	21	5	No_condition	Basic Audio Quality	1	
42	A9	3	Sys3	3	Sample3	1	No_condition	1	50	3	No_condition	Basic Audio Quality	4	
43	A9	3	Sys3	3	Sample3	1	No_condition	2	32	5	No_condition	Basic Audio Quality	5	
44	A9	4	Sys4	3	Sample3	1	No_condition	1	24	3	No_condition	Basic Audio Quality	8	
45	A9	4	Sys4	3	Sample3	1	No_condition	2	26	5	No_condition	Basic Audio Quality	3	
46	A9	5	Sys5	3	Sample3	1	No_condition	1	31	3	No_condition	Basic Audio Quality	5	
47	A9	5	Sys5	3	Sample3	1	No_condition	2	16	5	No_condition	Basic Audio Quality	7	
48	A9	6	Sys6	3	Sample3	1	No_condition	1	100	3	No_condition	Basic Audio Quality	6	
49	A9	6	Sys6	3	Sample3	1	No_condition	2	100	5	No_condition	Basic Audio Quality	6	

7 References

[1] G. Lorho, G. Le Ray, N. Zacharov, “eGauge – A Measure of Assessor Expertise in Audio Quality Evaluations” Proceeding of the Audio Engineering Society 38th International Conference on Sound Quality Evaluation, Piteå, Sweden, 13-15 June 2010.

[2] P.B. Brockhoff, Statistical testing of individual differences in sensory profiling. Food Quality and Preference 14(5-6), 425-434, 2003.

[3] ISO 8586-2, Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 2: Experts. International Organization for Standardization, 1994.

[4] G.B. Dijksterhuis and W.J. Heiser, The role of permutation tests in exploratory multivariate data analysis, Food quality and preference 6, 263-270, 1995.

[5] D.S. Moore, G.P. McCabe, Introduction to the Practice of Statistics, W.H. Freeman & Company, 2006.