Perceptual Audio Evaluation
Theory, method and application

Søren Bech
&
Nick Zacharov

What are we talking about?

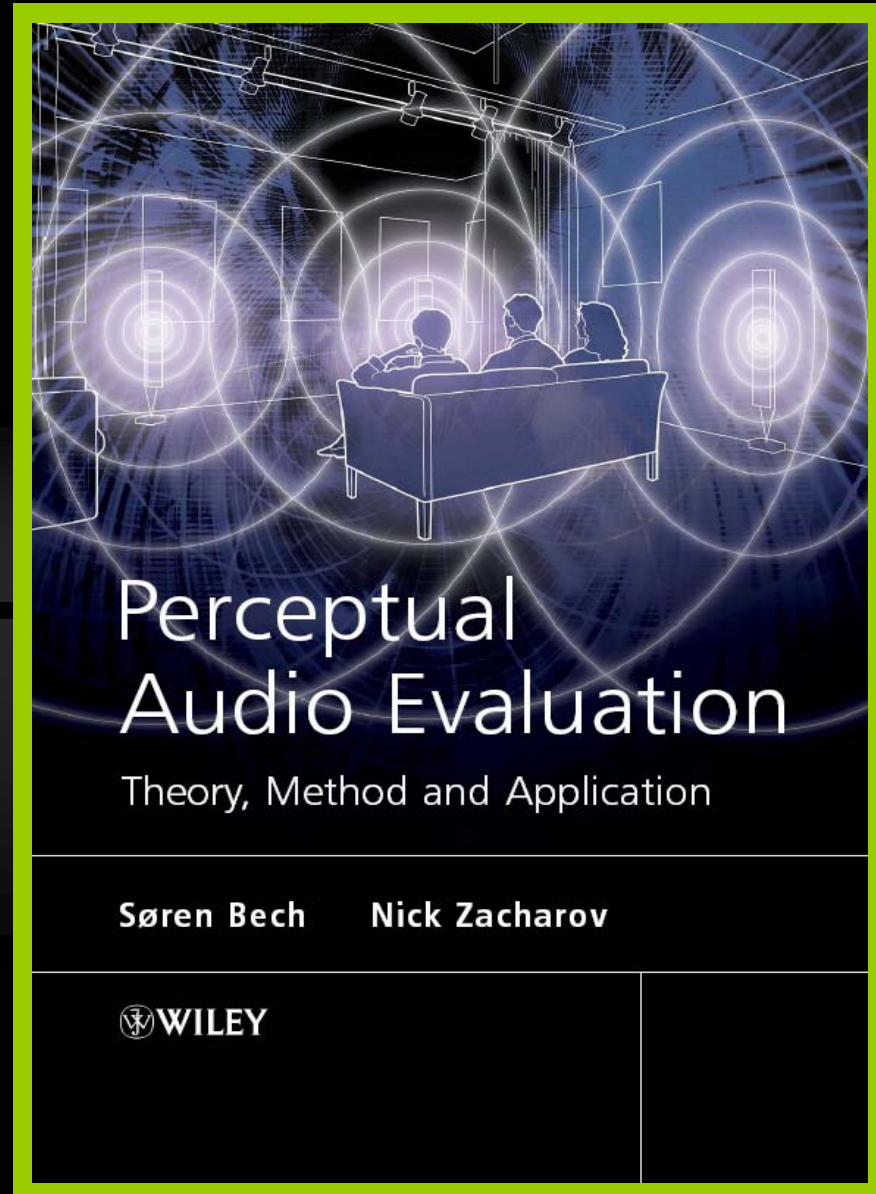**Søren Bech**

Bang & Olufsen

sbe@bang-olufsen.dk

Nick Zacharov

DELTA, SenseLab

NVZ@delta.dk

Perceptual
Audio Evaluation

Theory, Method and Application

Søren Bech   Nick Zacharov

WILEY

Introduction

This is a tutorial…

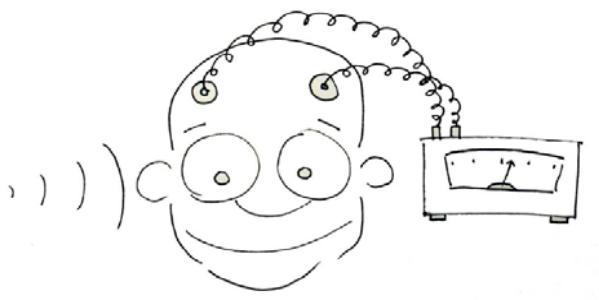…about listening tests…

…and how to perform them *well*

We will give you opportunity to ask questions as we go

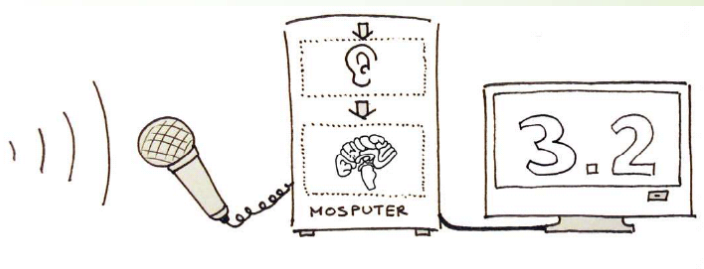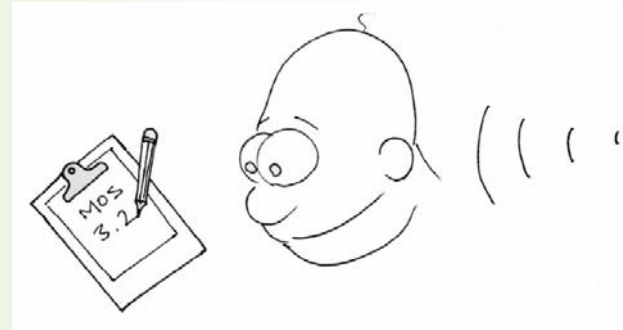And more time at the end of the session for general discussion

Audio quality assessment

Direct audio quality measurement not possible

Indirect audio quality measurement is practical
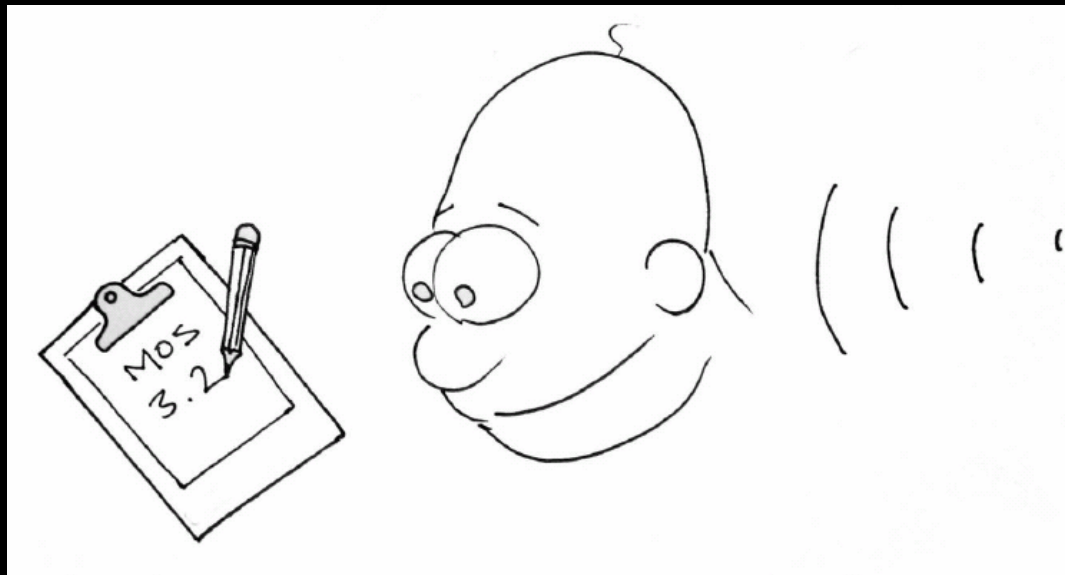
→ Listening tests

Audio quality prediction is developing

→ Predictions *only* as good as models

Our focus today…

Motivation for listening tests

✓ Listening test are needed when

- ✓ Physical measurement provide insufficient information
- ✓ Direct measurement of the perceived audio quality doesn't exist
- ✓ A predictive model of audio quality is not available

When to apply listening tests

To

- ✓ Study whether stimuli are perceptually identical
- ✓ Consider which sample is perceptually superior
    - ✓ and to what extent
- ✓ Establish which audio system is preferred
- ✓ Establish whether an audio system is acceptable for a given task
- ✓ Study the performance of audio systems in a detailed manner using perceptual attributes
- ✓ Define the absolute audio quality of an audio system

What listening test don't provide

- ✓ Identification of problematic system design parameters
- ✓ Finding the highest scoring system in a hifi magazine test
- ✓ Identifying what technical aspects makes a competitors solution superior

Types of listening tests

- **Perceptual measurement**
  - An objective quantification of the sensory strength of individual auditory attributes of the perceived stimulus

- **Affective measurement**
  - An objective quantification of an overall impression of the perceived stimulus

Distribution of this tutorial

✓	Introduction	→ Nick (done ☺)
✓	Definition of research question and hypothesis	→ Søren
✓	Fundamentals of experimentation	→ Søren
✓	Quantification of impression	→ Søren
✓	Statistics	→ Søren
✓	Experimental variable	→ Nick
✓	Technical considerations	→ Nick
✓	(Standards overview)	→ Nick

.....over to Søren....

Definition of research question and hypothesis

Definition of research question and hypothesis

Research question

General formulation of the problem to be investigated

Research question => hypothesis => design of experiment => conclusion => hypothesis true or false => research question answered

Definition of research question and hypothesis
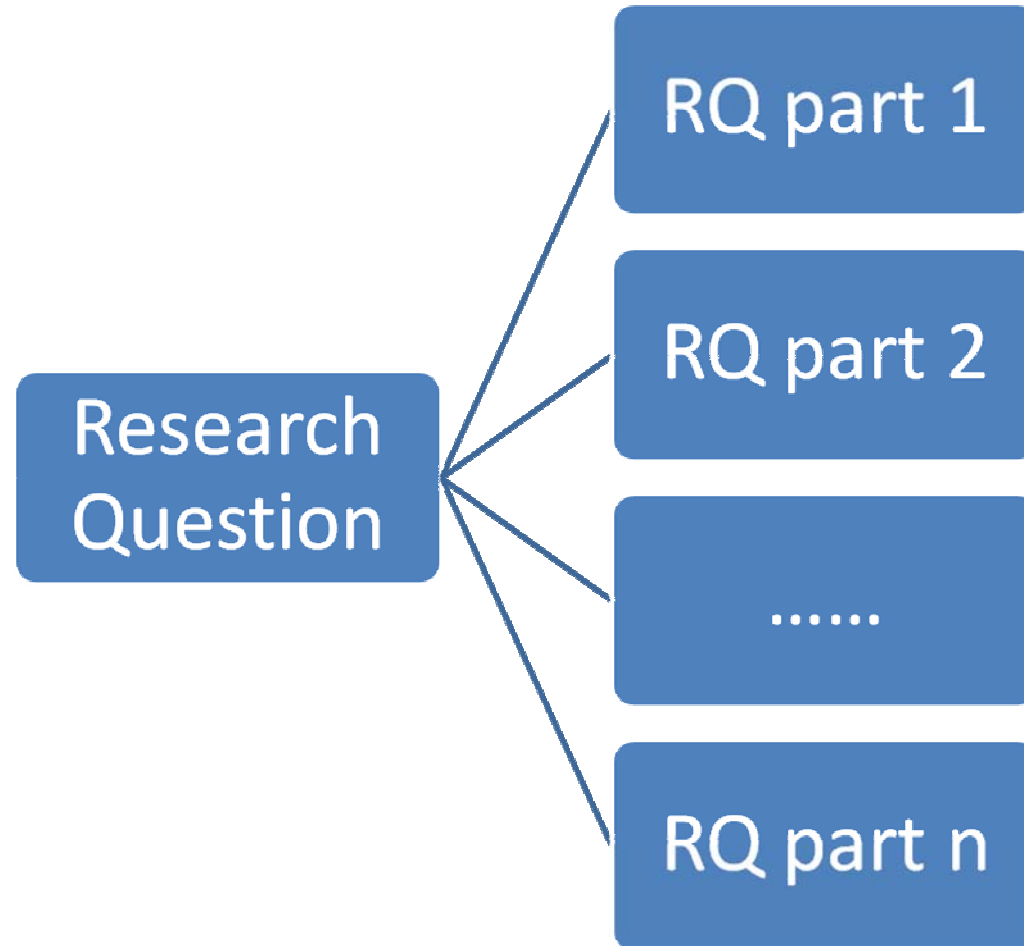
Definition of research question and hypothesis

Research Question

RQ part 1

RQ part 2

……
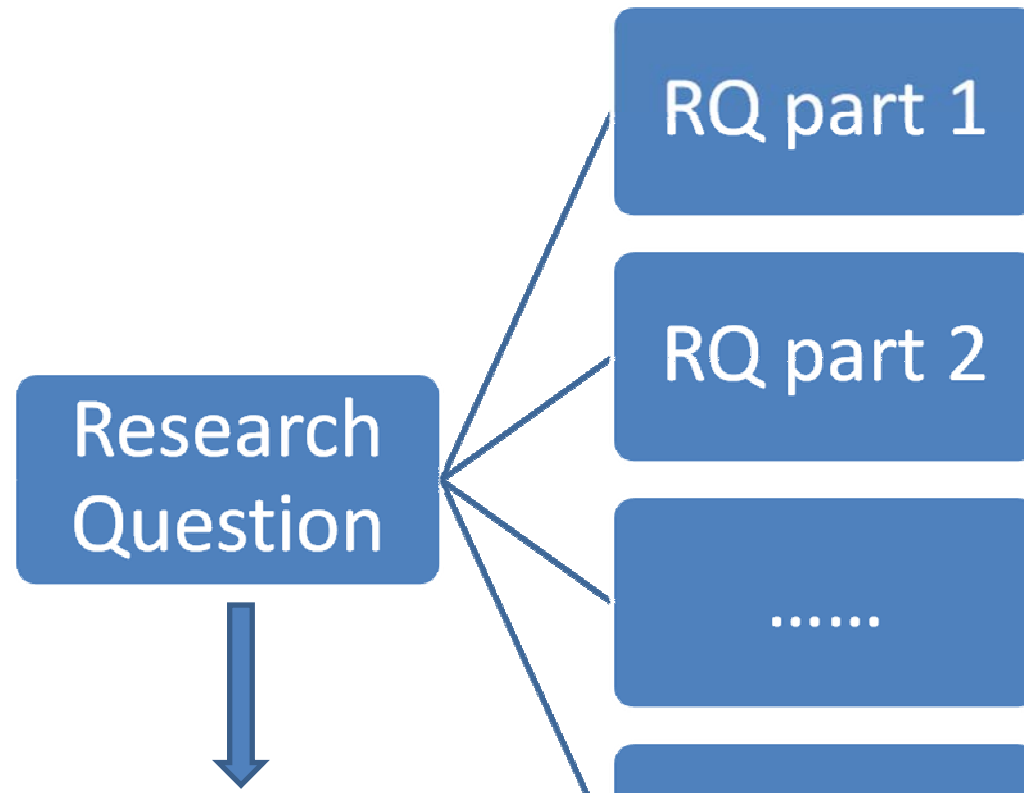
Why is the perceived sound quality of a loudspeaker different in different domestic sized rooms and in different positions in the same room? (Archimedes project)

Definition of research question and hypothesis

RQ part 1

RQ part 2

Do two physically identical loudspeakers positioned in two different positions in the same room sound different ?

Why is the perceived sound quality of a loudspeaker different in different domestic sized rooms and in different positions in the same room? (Archimedes project)

Definition of hypothesis

Research question part 1:

Do two, physically identical, loudspeakers positioned in two different positions in the same room sound different ?

To produce, scientifically valid, experimental verification of a hypothesis two principles can be applied:

- Empiricism
- Rationalism

Most experiments involving human subjects will be based on a combination of the two principles, but the principle of rationalism will (should) often be dominating.

Definition of hypothesis - rationalism

Premiss 1 (hypothesis)**:** It rains

Premiss 2 (initial conditions): When it rains the street will be wet

Conclusion (observable statement): The street will be wet

Rationalism: If the premises are true it follows that the conclusion will be true

Principle of verification:

Experimental observation: The streets are wet

=> Conclusion true => It rains

or a street cleaning truck just sprinkled water on the streets !

Principle of falsification:

Experimental observation: The street is not wet

Conclusion false => one of the premises are false => Its not raining ☺

Definition of hypothesis - rationalism

Research Question part 1: Do two physically identical loudspeakers positioned in two different positions in the same room sound different ?

Hypothesis (premiss 1): Physically identical loudspeakers positioned in different positions in the same room sound identical.

Initial conditions (premiss 2): Two physically identical loudspeakers in two different positions in the same room are compared in a listening test and the results are the **true** representations of the perceived sound quality of the individual loudspeakers

$\Rightarrow$ The initial conditions must be true

$\Rightarrow$ Careful experimental planning and control of all variables !

Fundamentals of experimentation

```
┌─────────────┐        ┌──────────────────┐        ┌──────────────┐
│ START       │        │ Definition of    │        │ Literature   │
│ Request for │───────▶│ research question│───────▶│ study        │
│ perceptual  │        │ and hypothesis   │        │              │
│ evaluation  │        └──────────────────┘        └──────────────┘
└─────────────┘
```

START Request for perceptual evaluation → Definition of research question and hypothesis → Literature study

Self evident conclusions - no evaluation needed

Very large

Suitable standard exists? ◀— No — Suitable perceptual model exists? ◀— Expected magnitude of perceptual diffs.? —Yes—▶ Is an evaluation needed ?

Yes — Follow procedure defined in standard

Yes — Use objective perceptual model

Very small — Too many observations needed - evaluation not feasible

No — Use results from literature

No

No

Define initial conditions → Statistical design and blocking → Establish statistical model → Conduct pilot experiment

Check design with dummy dataset

Document experimental setup

END

Report results

Yes

Results sufficient? ◀— Conclusions ◀— Statistical analysis ◀— Run (main) experiment

START Request for perceptual evaluation → Definition of research question and hypothesis → Literature study

Self evident conclusions - no evaluation needed

Very large

Suitable standard exist? — No → Suitable perceptual model exists? → Expected magnitude of perceptual effect → Is an evaluation needed ?

Many factors to consider

No

Yes

Follow procedure defined in standard

Yes

Use objective perceptual model

Very small

Too many observations needed - evaluation not feasible

Yes

No

Use results from literature

the tutorial ☺

No

Define initial conditions → Statistical design and blocking → Establish statistical model → Conduct pilot experiment

Check design with dummy dataset

Document experimental setup

END

Yes

Report results

Results sufficient? ← Conclusions ← Statistical analysis ← Run (main) experiment

Fundamentals of experimentation – filter model

Physical domain   **Perceptive domain**   **Affective domain**

Filter 1, the senses   Filter 2, non sound factors

Physical stimulus → **M1**   **Perceived stimulus** → **M2** +/-   **M3** **Likes/dislikes**

Sensory sensitivity and selectivity

Mood
Context
Emotion
Background
Expectation

Objective   Subjective

T. Holm Pedersen, 2001

2
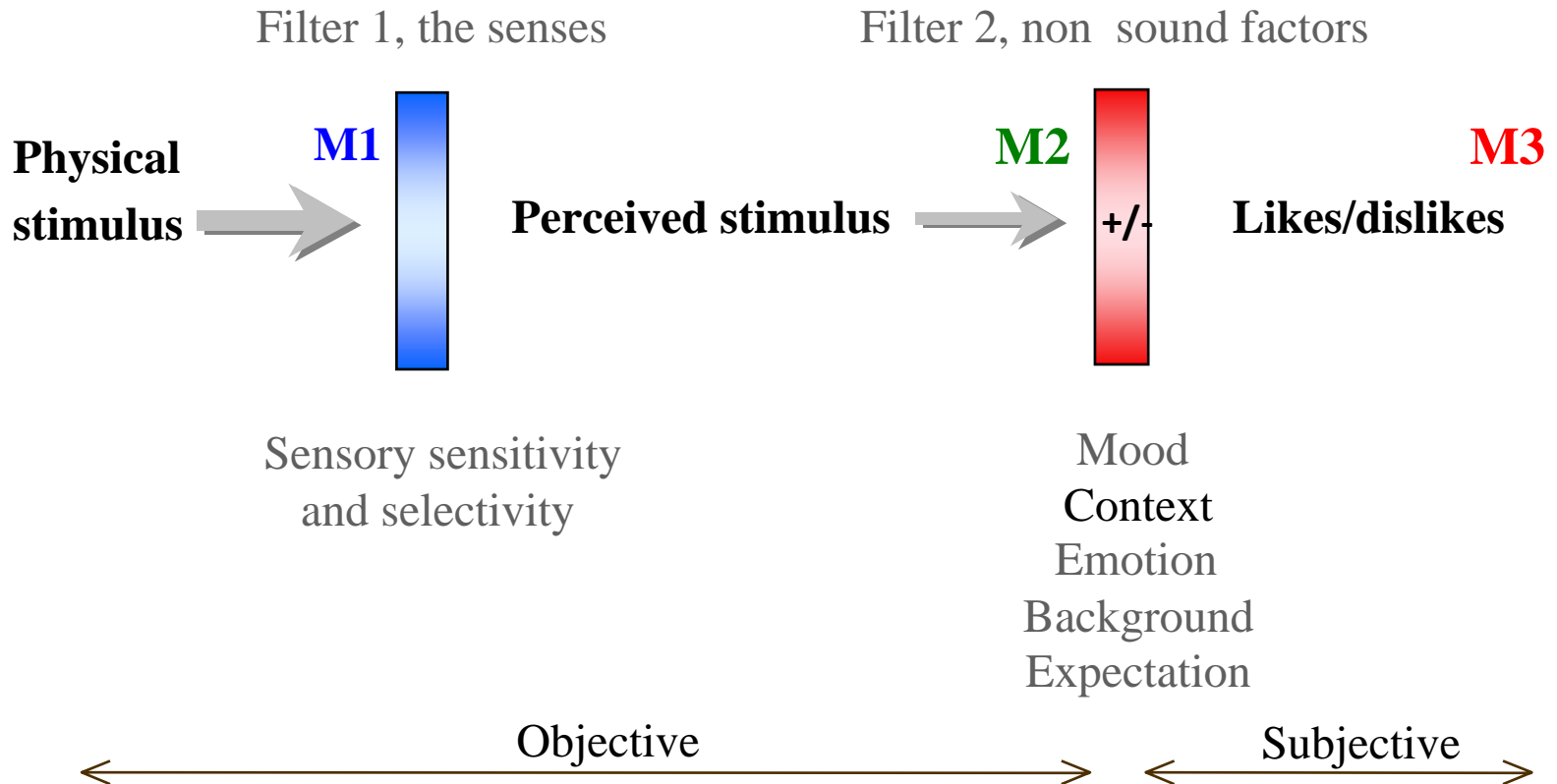
Fundamentals of experimentation – filter model

Physical domain

Measures

- Free-field frequency response
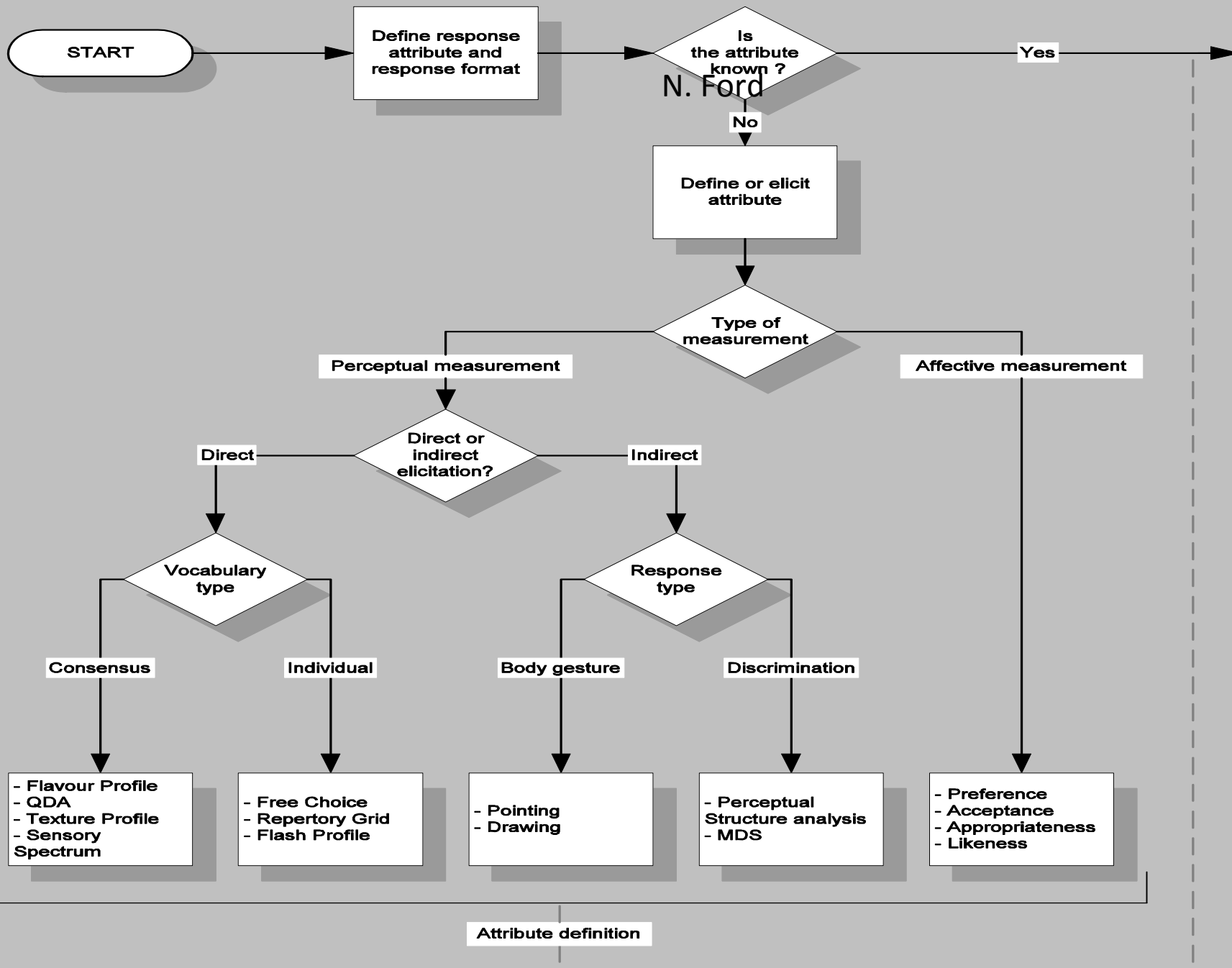- Max SPL
- Distortion
- ??

Perceptive domain

Attributes

- Loudness
- Spatial position
- Spatial blur
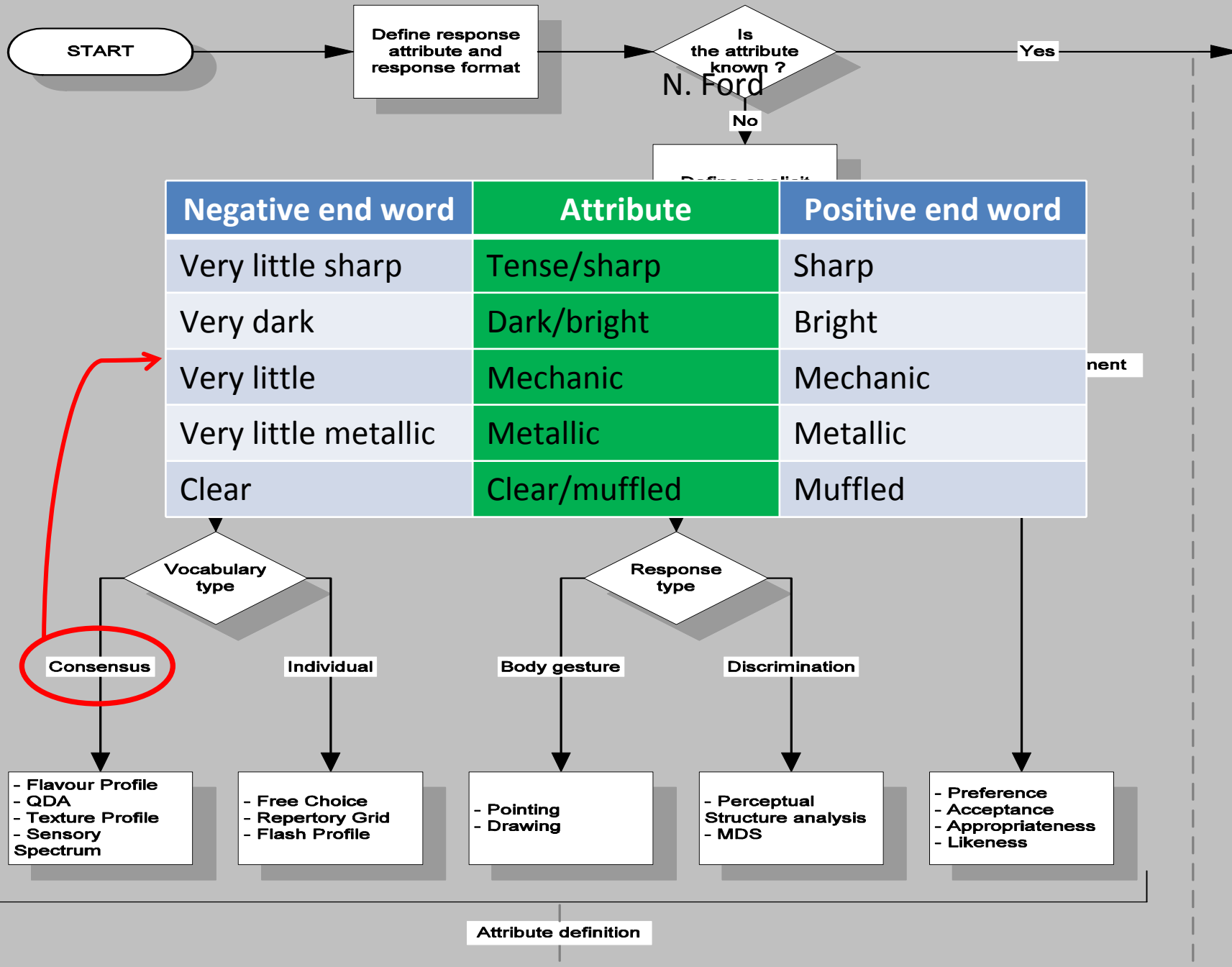- Timbre of lower bass
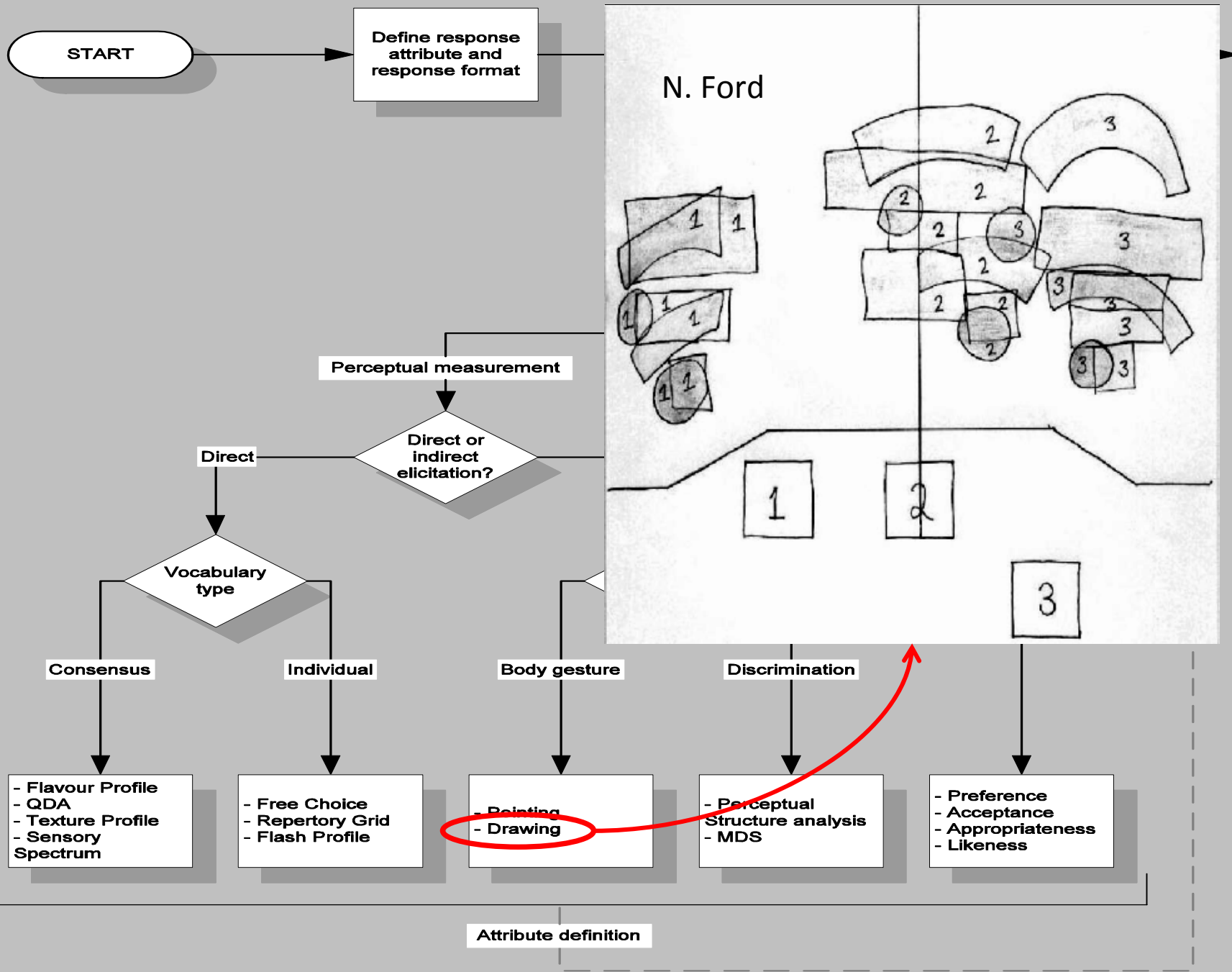- Timbre of midrange
- boxiness

Affective domain

Affective

- Like/dislike
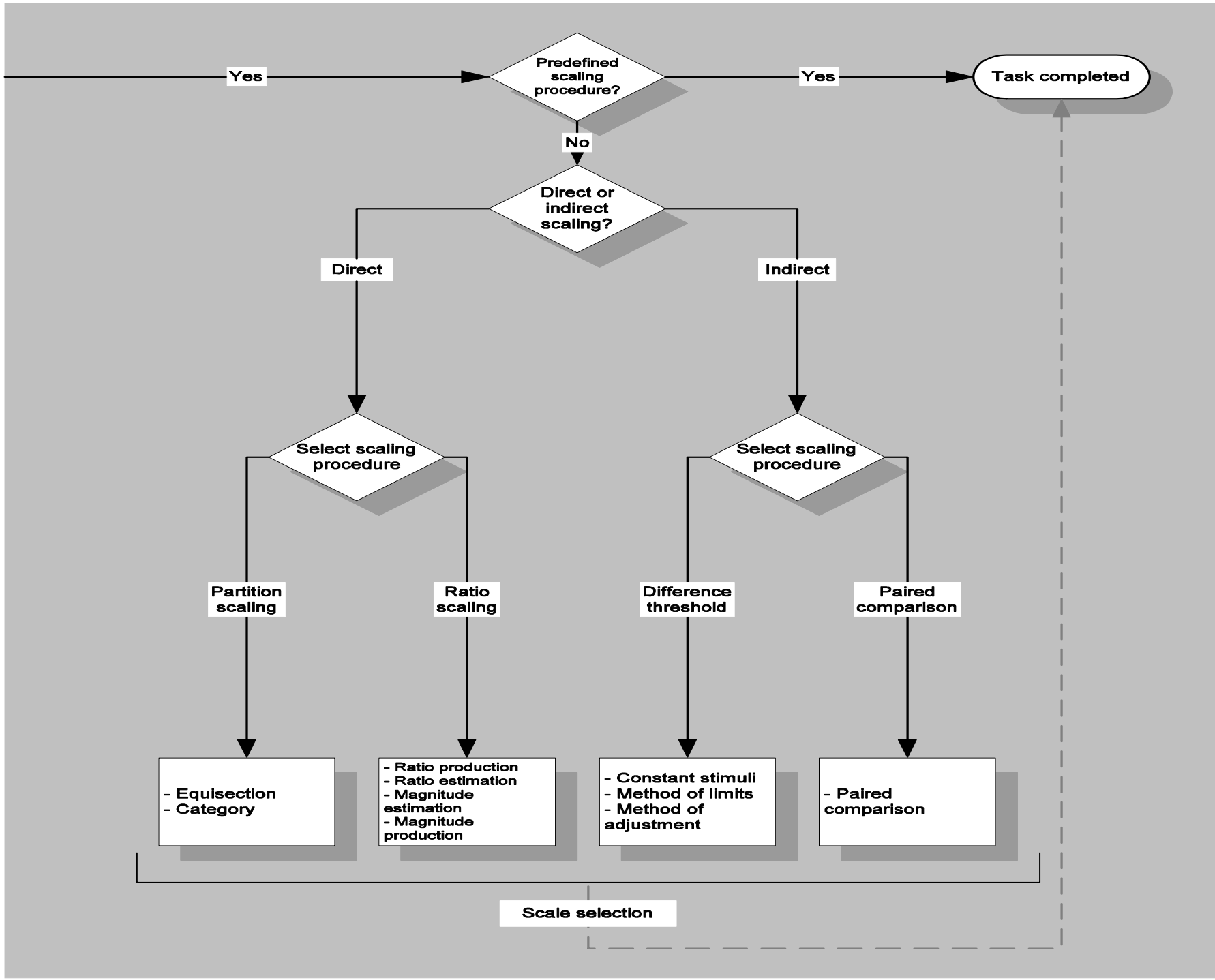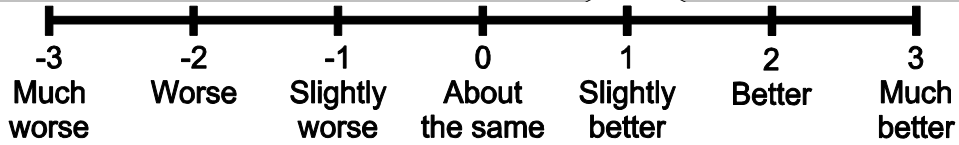- Acceptance
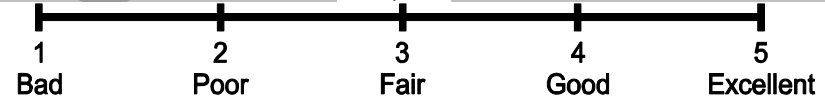- Annoyance

+/-

Quantification of impression

START → Define response attribute and response format → Is the attribute known ? — Yes

N. Ford

No ↓

Define or elicit attribute

↓

Type of measurement

Perceptual measurement | Affective measurement

Direct or indirect elicitation?

Direct — Indirect

Vocabulary type

Consensus | Individual

Response type

Body gesture | Discrimination

- Flavour Profile
- QDA
- Texture Profile
- Sensory Spectrum

- Free Choice
- Repertory Grid
- Flash Profile

- Pointing
- Drawing

- Perceptual Structure analysis
- MDS

- Preference
- Acceptance
- Appropriateness
- Likeness

Attribute definition

START → Define response attribute and response format → Is the attribute known ?

N. Ford

Yes

No

Define or elicit ... nent

Negative end word	Attribute	Positive end word
Very little sharp	Tense/sharp	Sharp
Very dark	Dark/bright	Bright
Very little	Mechanic	Mechanic
Very little metallic	Metallic	Metallic
Clear	Clear/muffled	Muffled

Vocabulary type

Response type

Consensus

Individual

Body gesture

Discrimination

- Flavour Profile
- QDA
- Texture Profile
- Sensory Spectrum

- Free Choice
- Repertory Grid
- Flash Profile

- Pointing
- Drawing

- Perceptual Structure analysis
- MDS

- Preference
- Acceptance
- Appropriateness
- Likeness

Attribute definition

START

Define response attribute and response format

N. Ford

Perceptual measurement

Direct or indirect elicitation?

Direct

Vocabulary type

Consensus

Individual

Body gesture

Discrimination

- Flavour Profile
- QDA
- Texture Profile
- Sensory Spectrum

- Free Choice
- Repertory Grid
- Flash Profile

- Pointing
- Drawing

- Perceptual Structure analysis
- MDS

- Preference
- Acceptance
- Appropriateness
- Likeness

Attribute definition

```
                                          ┌─────────────┐
  ─── Yes ──────────────►  Predefined  ─── Yes ──────►  Task completed
                          scaling
                          procedure?

                            No
                            ▼

                          Direct or
                          indirect
                          scaling?

         Direct                              Indirect

           ▼                                    ▼

      Select scaling                       Select scaling
      procedure                            procedure

  Partition          Ratio          Difference        Paired
  scaling            scaling        threshold         comparison

      ▼                ▼                ▼                ▼

 - Equisection   - Ratio production  - Constant stimuli  - Paired
 - Category      - Ratio estimation  - Method of limits    comparison
                 - Magnitude         - Method of
                   estimation          adjustment
                 - Magnitude
                   production
```

Scale selection

Comparison category rating (CCR) scale

-3	-2	-1	0	1	2	3
Much worse	Worse	Slightly worse	About the same	Slightly better	Better	Much better

Task completed

ct or indirect scaling?

Direct

ITU-T P.800 absolute category rating scale

1	2	3	4	5
Bad	Poor	Fair	Good	Excellent

ITU-R 5-point continuous impairment scale

1	2	3	4	5
Very annoying	Annoying	Slightly annoying	Perceptible, but not annoying	Imperceptible

caling dure

Partition scaling

Ratio scaling

Difference threshold

Paired comparison

9-point hedonic categorical scale

-4	-3	-2	-1	0	1	2	3	4
Dislike extremely	Dislike very much	Dislike moderately	Dislike slightly	Neither like nor dislike	Like slightly	Like moderately	Like very much	Like extremely

- Equisection
- Category

ired parison

production

Scale selection

Statistical considerations

Major statistical issues

- Statistical design of experiment
- Testing of statistical assumptions
- Analysis of data

This talk will focus on the design of experiments as this determines the basic quality of the results.

For the two others:

consult a statistician – or the book ☺

The basic statistical question

Is the observed variability in the subjective impression a result of the presented stimuli (e.g. loudspeaker–programme combinations) or is it random fluctuations ?

YES: There are, statistically significant, audible differences between some of the presented loudspeaker-programme combinations

NO: There are no, statistically significant, audible differences between the loudspeaker-programme combinations

A statistical model

$$Y_{t,i} = \mu + \alpha_t + \varepsilon_{t,i}$$

$Y_{t,i}$ is the i'th rating provided by a subject for the t'th stimuli

μ is the general mean for all stimuli

α_t is the effect caused by the t'th stimuli when corrected for the overall mean

$\varepsilon_{t,i}$ is the effect caused by the random experimental error

Experimental variables

- Explanatory variables
  - Dependent variables ($Y_{t,i}$)
  - Independent variables (α_t)
- Controlled variables (μ , $\varepsilon_{t,i}$)
- Disturbing variables (μ , α_t , $\varepsilon_{t,i}$)
- Randomised variables ($\varepsilon_{t,i}$)

$$ Y_{t,i} = \mu + \alpha_t + \varepsilon_{t,i} $$

Experimental variables

- Dependent variables ($Y_{t,i}$)
  - Variables quantifying the subjects impression(s) (timbre, spatial impression etc) of the presented stimuli

- Independent variables (α_t)
  - Variables under investigation and controlled by the experimenter (loudspeakers, programmes, SPL, base angle etc.)

- Controlled variables (μ , $\varepsilon_{t,i}$)
  - Variables known to the experimenter, but not a part of the research question – can be included in the statistical model as separate elements (room position of the loudspeaker, listening room etc.)

Experimental variables

- Disturbing variables (μ , α_t , $\varepsilon_{t,i}$)
  - Variables unknown to the experimenter. The experimental design aims at "converting" these into random variables,

- Randomised variables ($\varepsilon_{t,i}$)
  - Variables unknown to the experimenter, but that by nature will influence the dependent variable in a completely random manner

Experimental design

The purpose of the experimental design is to control the influence of the disturbing variables such that they do not influence the dependent, independent and controlled variables

- Treatment design
  – Specification of which stimuli to use and how to administer these independently of the subjects,
- Allocation of stimuli design
  – Specification of how to administer the stimuli to the individual subjects

Treatment design

- Full factorial design
  - The presented stimuli represent all possible combinations of the independent variables

- Fractional design
  - The presented stimuli represents a subset of all possible combinations of the independent variables

Treatment design – an example

Definition of experiment

- **Hypothesis:** Physically identical loudspeakers sound identical in the same position in different rooms and in different positions in the same room
- **Dependent var.:** fidelity of timbre of reproduction
- **Independent vars.:** listening room (4), loudspeaker position in room (4)
- **Controlled var.:** loudspeakers (4 levels), programme (1)

=> Full factorial = 64 (4x4x4x1) stimuli

But: rooms must be used in parallel and only one position per loudspeaker per room

=> Fractional design using 4 identical copies of each loudspeaker

Treatment design – an example

- Four loudspeakers (A, B, C, D)
- Four rooms (R1, R2, R3, R4)
- Four loudspeaker positions (P1, P2, P3, P4)

Des. I	R1	R2	R3	R4	average
P1	A	B	C	D	Y1
P2	A	B	C	D	Y2
P3	A	B	C	D	Y3
P4	A	B	C	D	Y4
average	x1	x2	x3	x4	

Treatment design – an example

Problem with design I: Differences between room mean ratings could be due to differences between loudspeakers and differences between listening rooms ☹

=> Loudspeaker and room variables are confounded

Des. I	R1	R2	R3	R4	average
P1	A	B	C	D	Y1
P2	A	B	C	D	Y2
P3	A	B	C	D	Y3
P4	A	B	C	D	Y4
average	x1	x2	x3	x4	

Treatment design – an example

- Four loudspeakers (A, B, C, D)
- Four rooms (R1, R2, R3, R4)
- Four loudspeaker positions (P1, P2, P3, P4)

Des. II	R1	R2	R3	R4	average
P1	A	A	A	A	Y1
P2	B	B	B	B	Y2
P3	C	C	C	C	Y3
P4	D	D	D	D	Y4
average	x1	x2	x3	x4	

Treatment design – an example

Problem with design II: Differences between position mean ratings could be due to differences between loudspeakers and differences between positions in listening rooms ☹

=> **Loudspeaker and position variables are confounded**

Des. II	R1	R2	R3	R4	average
P1	A	A	A	A	Y1
P2	B	B	B	B	Y2
P3	C	C	C	C	Y3
P4	D	D	D	D	Y4
average	x1	x2	x3	x4	

Treatment design – an example

Randomised design

Des. III	R1	R2	R3	R4
P1	C	A	D	A
P2	A	A	C	D
P3	D	B	B	B
P4	D	C	B	C

Treatment design – an example

Problem with design III: Loudspeakers A & B are only tested in three of the four rooms and in two of the four positions ☹

=> **Loudspeaker, position and room variables are confounded**

Des. III	R1	R2	R3	R4
P1	C	A	D	A
P2	A	A	C	D
P3	D	B	B	B
P4	D	C	B	C

Treatment design – an example

Randomised block (room) design

Des. IV	R1	R2	R3	R4
P1	B	D	A	C
P2	C	C	B	D
P3	A	A	D	B
P4	D	B	C	A

Treatment design – an example

Problem with design IV: All loudspeakers are now tested in all four rooms, but loudspeakers A & C are only tested in three of the four positions ☹

=> Loudspeaker and position variables are confounded

Des. IV	R1	R2	R3	R4
P1	B	D	A	C
P2	C	C	B	D
P3	A	A	D	B
P4	D	B	C	A

Treatment design – an example

Latin Square design

Des. V	R1	R2	R3	R4
P1	B	D	A	C
P2	C	A	B	D
P3	A	C	D	B
P4	D	B	C	A

Treatment design – an example

Row means => test influence of position ☺

Column means => test influence of room ☺

Des. V	R1	R2	R3	R4
P1	B	D	A	C
P2	C	A	B	D
P3	A	C	D	B
P4	D	B	C	A

Allocation of stimuli design

- ## Within-subject:
  - Each subject receives all stimuli specified by the treatment design => All subjects visit all listening rooms

- ## Between-subject:
  - Each subject receives only one stimulus of those specified by the treatment design => one subject only visits one listening room

Allocation of stimuli design

Within-subjects designs are preferable to between-subjects design because:

- More efficient use of subject's time as more data is collected per visit
- Differences between subjects apply to entire stimuli set => more statistical "control"
- Smaller random error term as it only includes variation from the disturbing variables and not variations due to differences between subjects

Within-subjects designs

- Number of stimuli is often large and subjects can only work for 20 – 30 min.
  - Use fractional designs

- Stimulus presentation order needs to be controlled to avoid bias effects
  - Use Balanced Latin Square Designs

BLS for within-subjects design

Balanced Latin Square design for loudspeaker presentation order in room 1 and position 1

	Presentation number			
Subject number	1	2	3	4
1	A	B	D	C
2	B	C	A	D
3	C	D	B	A
4	D	A	C	B

BLS for within-subjects design

All loudspeakers are evaluated by all subjects and across subjects - each loudspeaker is presented after all other loudspeakers once and only once => presentation order effects is eliminated for averages across subjects

Subject number	1	2	3	4
1	A	B	D	C
2	B	C	A	D
3	C	D	B	A
4	D	A	C	B

Allocation of stimuli design

Basic question for both within- and between- subjects designs:

How many observations are needed pr stimulus ?

 => How many subjects and replications ?

Allocation of stimuli design

How many subjects and replications?

- What is the smallest subjective difference that needs to be resolved ?

- What is the variance of ratings per subjects ?

- What is the needed probability levels of Type I and Types II errors ?

Conduct a number of small (e.g. limited number of subjects and stimuli) pilot tests to estimate factors

Allocation of stimuli design

Consult a statistician – or the book ☺

(number of subjects and stimuli) pilot tests to estimate factors

over to Nick

Experimental variables

Experimental variables

✓ For robust experiments, independent experimental variables should be carefully controlled

✓ These include
- ✓ Signals
- ✓ Reproduction systems
- ✓ Reproduction or listening room (discussed later)
- ✓ Assessors
- ✓ (Systems under test)

Signals

- ✓ Test signals or program material
  - ✓ Must be well motivated
    - ✓ They should be
      - ✓ Representative
      - ✓ Challenging

  - ✓ Do not underestimate the selection process

  - ✓ With reference stimuli, ensure these are well motivated
    - ✓ Its not always obvious what the reference should be

  - ✓ Pilot experiments and expert listening is very valuable here

  - → Refer to the book for details

Reproduction systems

- ✓ Ensure the recording technique is compatible with the reproduction system
  - ✓ Apply appropriate transforms where needed

- ✓ Ensure that reproduction system is sufficient
  - ✓ Distortion
  - ✓ Level
  - ✓ Bandwidth

  - ✓ Refer to the book for more details

Assessors & listening panels

✓ Two types of panel

✓Consumer / Naïve assessor
  ✓Representative of consumers
  ✓Untrained
  ✓Selected based on demographic requirements or random sampling
  ✓Large panels (32…80…more)
  ✓Low investment

  →Affective testing

✓Expert assessor
  ✓Selected and trained
    ✓Population
    ✓Acuity
    ✓Ability
    ✓Availability
  ✓Small panels (N = 10 – 20)
  ✓High investment

  ✓Effective or descriptive testing

Assessor categorisation: ISO 8586-2

Assessor category	Definition
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
Expert	In the general sense, a person who through knowledge or experience has competence to give an opinion in the field about which he/she is consulted
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
Specialised expert assessor	Expert assessor who has additional experience as a specialist in the product and/or process and/or marketing, and who is able to perform sensory analysis of the product and to evaluate or predict effects of variations relating to raw materials, recipes, processing, storage, ageing, etc.

Assessor development
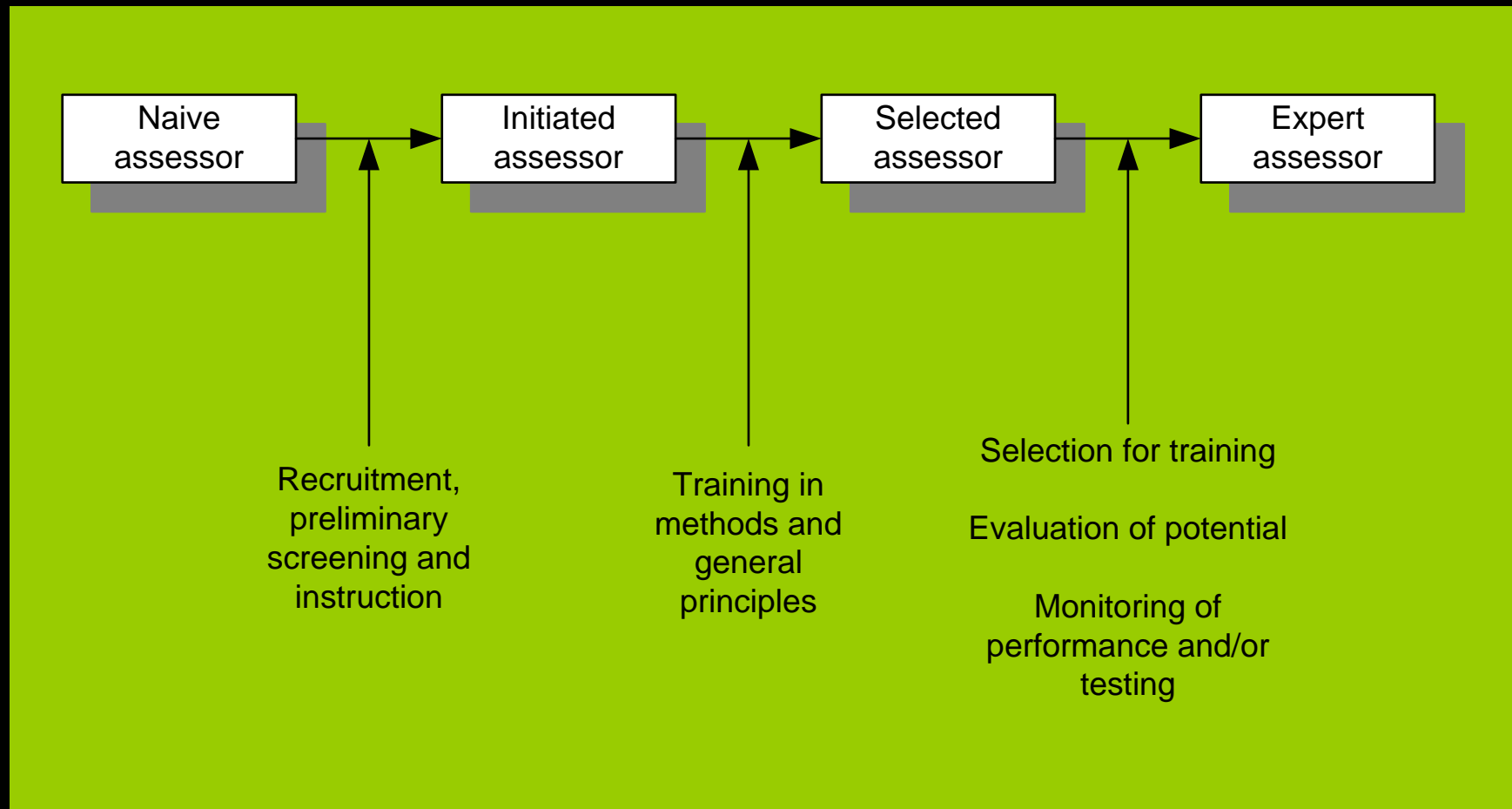
Panel selection procedure

✓ Questionaire

   ✓Evaluate whether subject are from the required population

✓ Audiometry

   ✓Check whether subjects have good physiology (Acuity)

✓ Screening tests

   ✓Evaluate if subjects are able discriminate and repeatably rate stimuli (Ability)
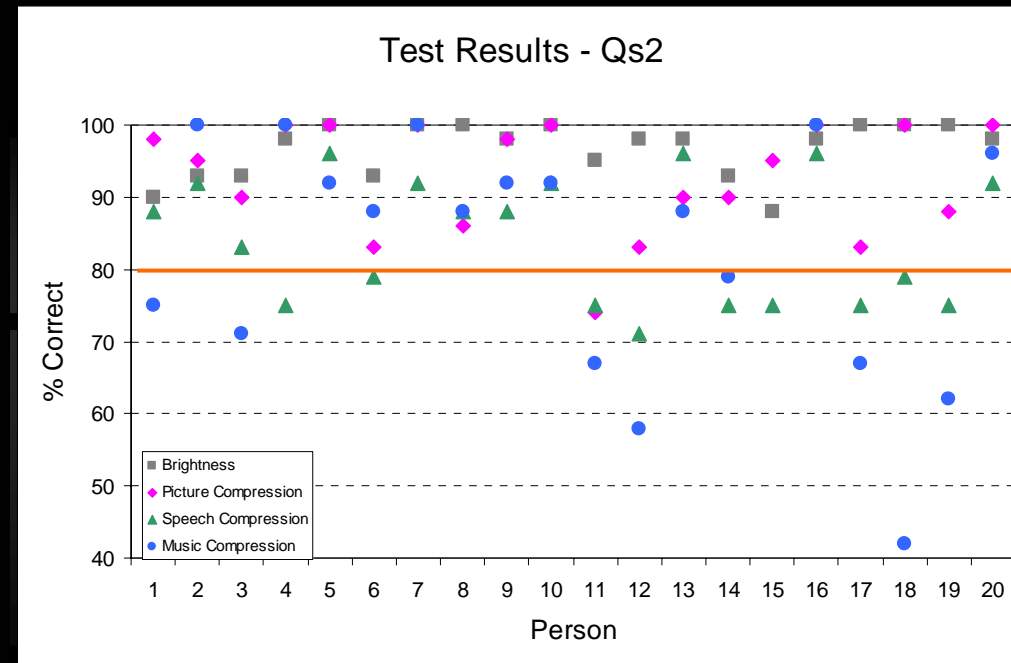
Expert assessor characteristics (1)

✓ The key to an assessors objectivity lies in the following 3 characteristics:

- ✓ Repeatability
  - ✓ The *precision* with which a subject can provide independent repeated ratings of the same test item
- ✓ Agreement
  - ✓ The level of agreement between a subject and the panel
- ✓ Discrimination
  - ✓ The ability to identify and rate perceptual difference between stimuli on attribute scales

✓ It is the evaluation of these characteristics which are important in the assessment of assessor/panel performance

Expert assessor characteristics (2)
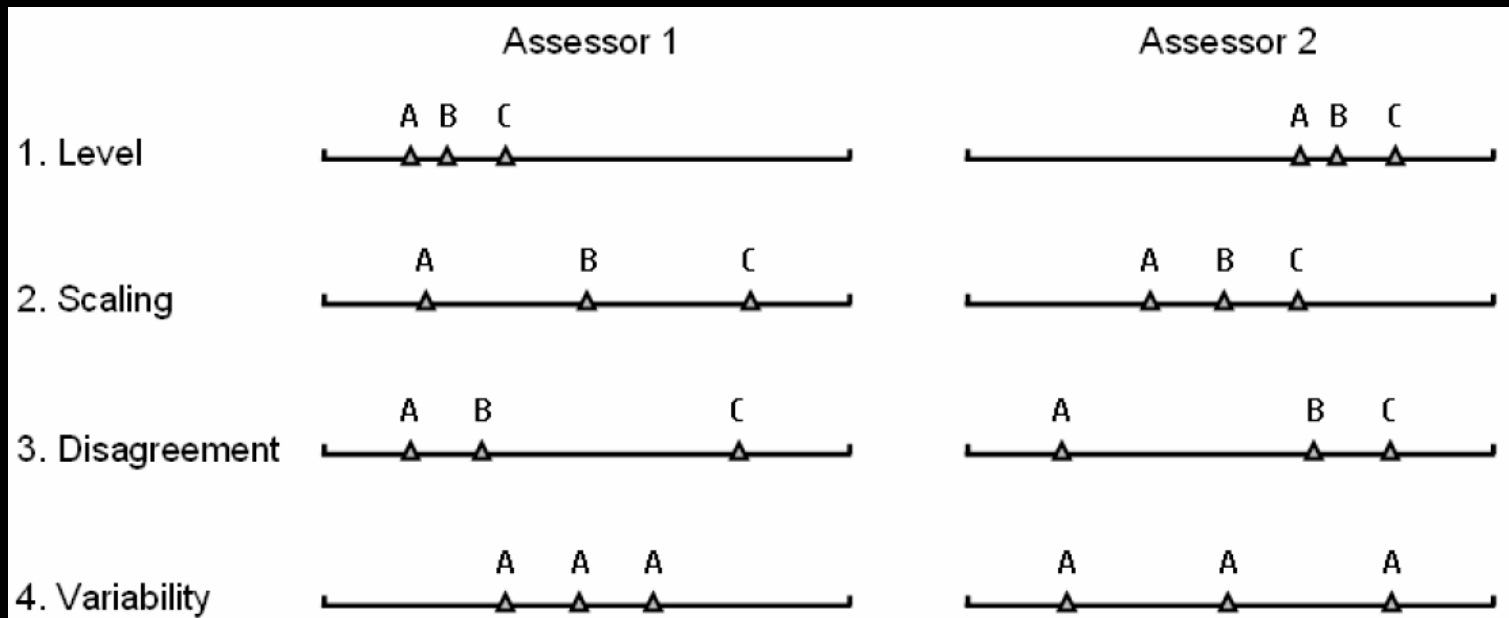
Assessor performance

- ✓ Six performance metrics

  - ✓ **Location**
    - ✓ The average score given by assessors
  - ✓ **Span**
    - ✓ The average standard deviation of a score given by an assessor within a session
  - ✓ **Unreliability**
    - ✓ Assessor unreliability based upon a measure of the session error
  - ✓ **Drift-Mood**
    - ✓ A measure of the between-sessions error that can be associated with assessor mood, and so on
  - ✓ **Discrimination**
    - ✓ A measure of an assessor's discrimination skills based upon the classical F-ratio for testing the significance of the product effect for an assessor
  - ✓ **Disagreement**
    - ✓ A measure of an assessor's disagreement based upon the assessor's contribution to Product * Assessor interaction F-ratio.

Measuring panel performance
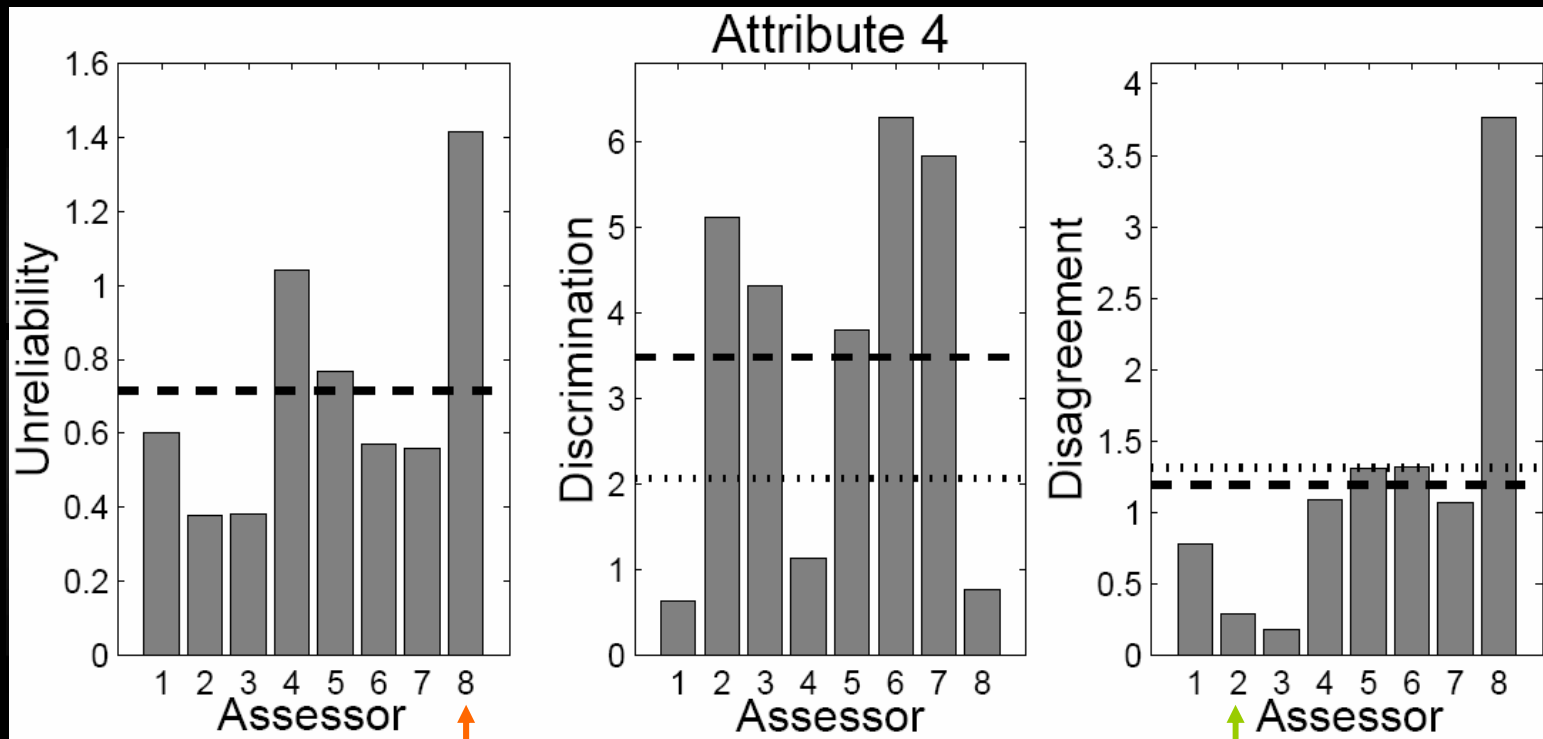
Technical considerations

Listening spaces

- ✓ Provide stable listening (& viewing) conditions
- ✓ *Somehow* representative of domestic listening spaces

- ✓ Main standards
  - ✓ ITU-R BS.1116-1
  - ✓ IEC 60268-13
  - ✓ EBU 3276

- ✓ Control of key factors
  - ✓ Reverberation
    - ✓ Defined reflection characteristics
  - ✓ Background noise
  - ✓ Sound & vibration isolation

  - ✓ Loudspeaker setups defined
  - ✓ Other factors
    - ✓ Ventilation
    - ✓ Lighting
    - ✓ Furnishing
    - ✓ Subject comfort
    - ✓ PC connectivity, etc...

ITU-R BS.1116-1 listening rooms

Loudspeaker setups

b = 2 - 4 m
h ≈ 0.9 b
θ ≈ 60°
r_L ≤ 0.7 m

Listening area
(radius r_L)

Display

b = 2 - 4 m
h ≈ 0.9 b
θ ≈ 60°
φ = 110° - 120°
r_L ≤ 0.8 m
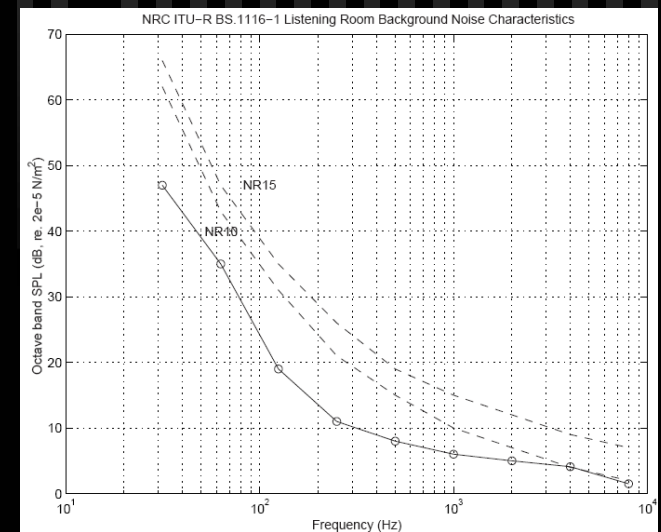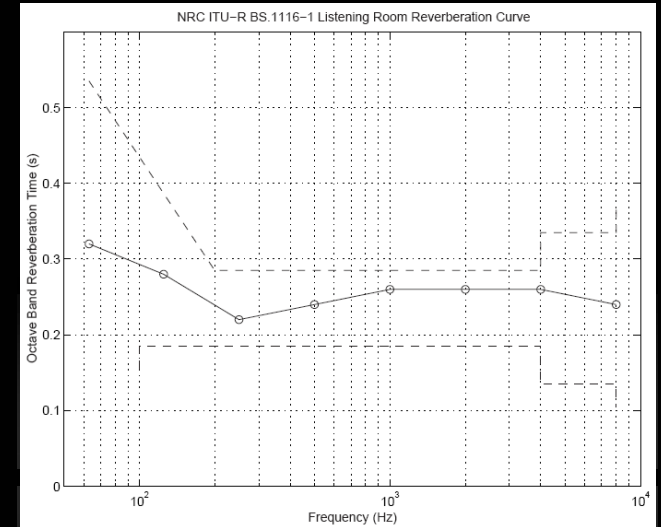
Listening area
(radius r_L)
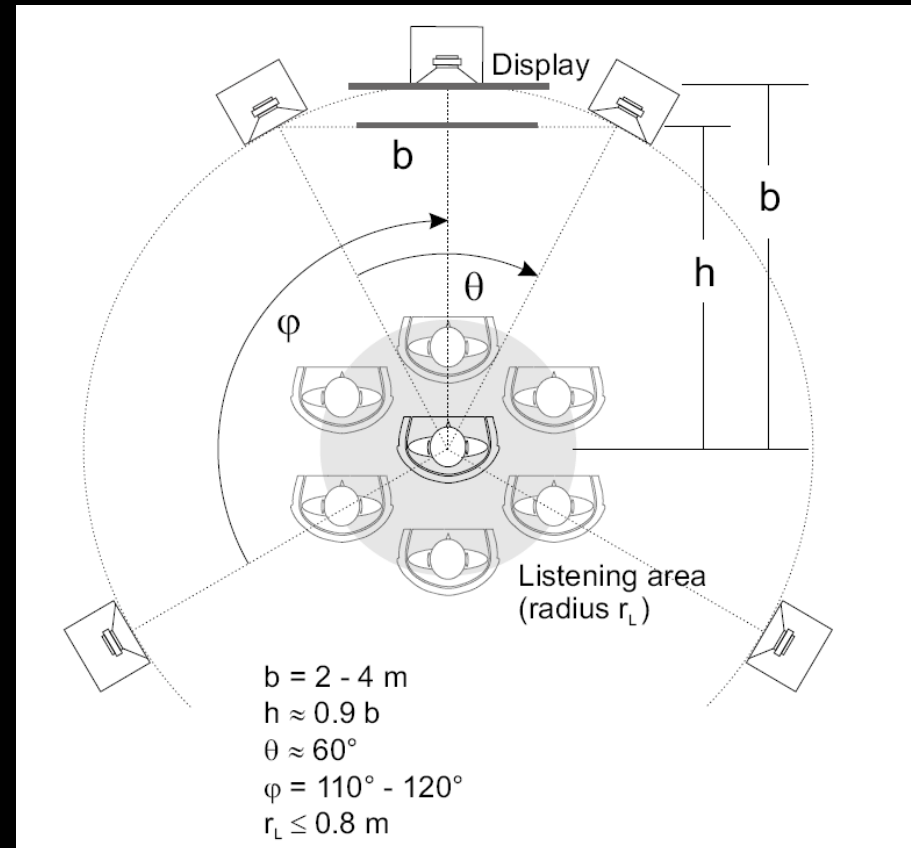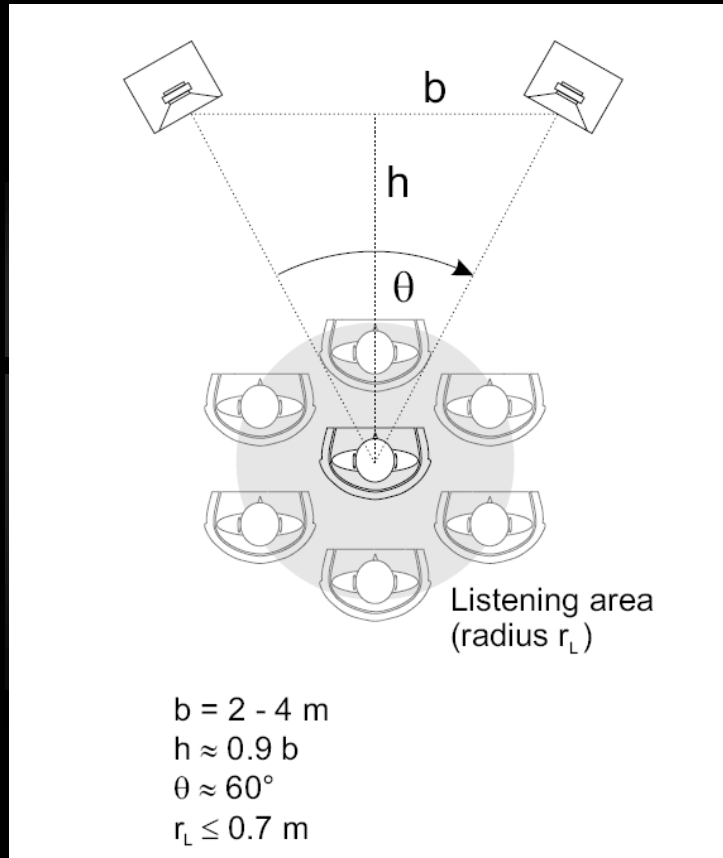
Common problems

- ✓ Mismatch between viewing and listening requirements
  - ✓ Check viewing/listening distances for you experiment early on
- ✓ Subwoofer calibration
  - ✓ This is still quite an open issue
- ✓ Electrical issues
  - ✓ Know your sound cards, DACs, amplifiers
    - ✓ Are they giving you the required level of performance (noise, distortion, etc.)

Level calibration

- ✓ Several different aspects
  - ✓ Absolute level calibration
    - ✓ Make sure you are within same limits
  - ✓ Inter-stimulus calibration
    - ✓ Ensure samples are at the same reproduction level
  - ✓ Inter-system calibration
    - ✓ Comparable loudness between reproduction systems
  - ✓ Inter-channel calibration
    - ✓ Mostly when using multichannel sound systems

  - ✓ Choose the most effective / suitable approach
    - ✓ Loudness, temporal loudness, dBA, etc....

Test planning checklist

- ✓ Experimental design and planning
- ✓ Subject selection
- ✓ Sample selection and processing
- ✓ Configuration and calibration of the experimental setup
- ✓ Pilot study
  - ✓ Analysis, refinement, etc…
- ✓ Main study
- ✓ Analysis
- ✓ Reporting

Logistics

- ✓ WHERE
    - ✓ Is there a suitable room for the test?
    - ✓ Does it meet the technical requirements?
- ✓ HOW
    - ✓ How shall the test be administered?
    - ✓ How will subjects be trained, instructed and escorted
- ✓ WHO
    - ✓ Who is performing the test? Does a panel exist and are they suitably trained?
- ✓ WHAT
    - ✓ Is all the required equipment available?
- ✓ WHEN
    - ✓ Book rooms, facilities and subjects for the setup and testing periods

Listening test administration
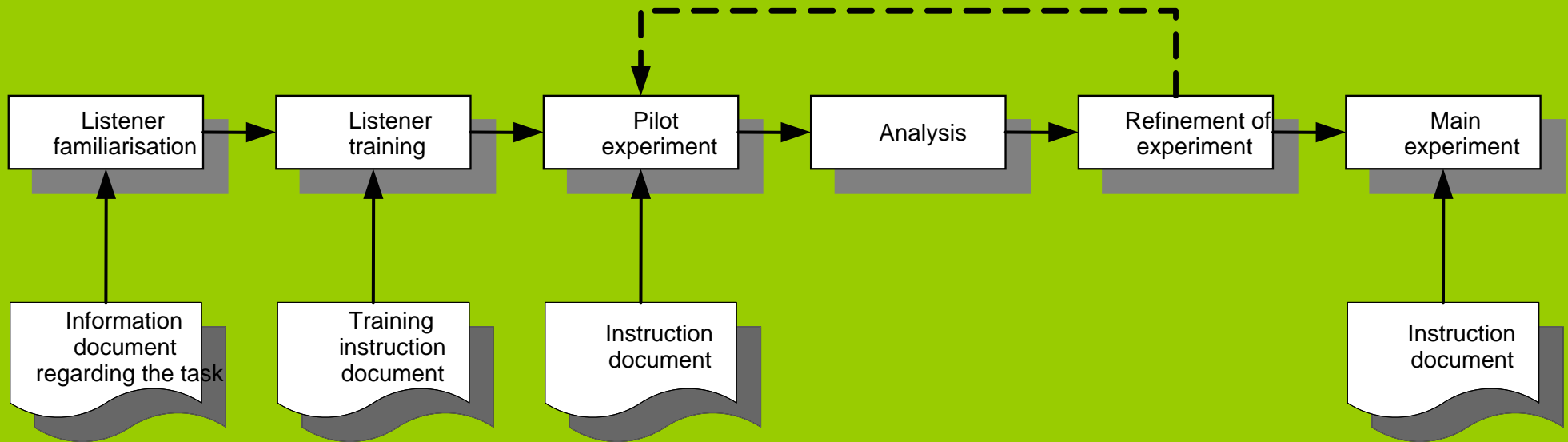
Ethical consideration

Beware of
- ✓ Data security & privacy of information
- ✓ Listening levels
  - ✓ Avoid harm !
- ✓ General ethical considerations
  - ✓ Unfair discrimination
  - ✓ Sexual harassment
  - ✓ Other harassment
  - ✓ Avoiding harm
  - ✓ Multiple relationships
  - ✓ Conflict of interest
  - ✓ Third-party requests for services
  - ✓ Exploitative relationships
  - ✓ Cooperation with other professionals
  - ✓ Informed consent

Listening test software

Some example tools

- ✓ STEP
  - ✓ Windows based
  - ✓ ITU-R BS.1116-1, ITU-R BS.1534-1, ITU-T P.800 ACR
- ✓ CRC-SEAQ
  - ✓ Windows based
  - ✓ ITU-R BS.1116-1, ITU-R BS.1534-1
- ✓ Fraunhofer MUSHRA Software
  - ✓ Windows/Unix
  - ✓ ITU-R BS.1534-1 MUSHRA tests
- ✓ PCABX
  - ✓ Window ABX testing
- ✓ MUSHRAM
  - ✓ Matlab GUI for ITU-R BS.1534-1 MUSHRA tests

Standards
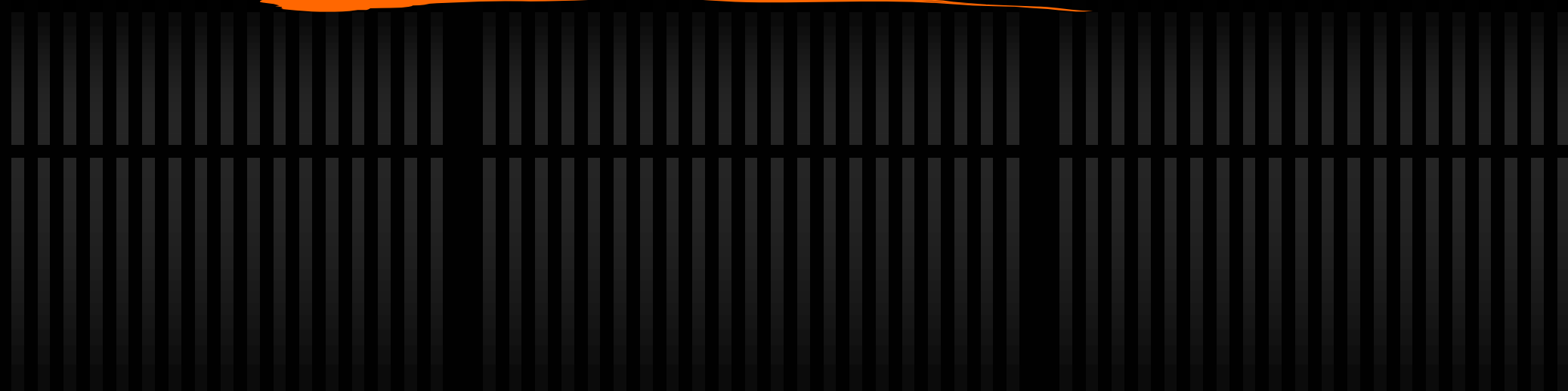
The role of Standards

- ✓ Over 30 listening test related standards
- ✓ Provide agreed best practices
  - ✓ For specific applications
- ✓ Mostly based on overall quality measures
  - ✓ Mean Opinion Score (MOS) or similar
- ✓ Advances methods often outside the scope of standards

- ✓ Key organisations
  - ✓ ITU-R, ITU-T, AES, IEC,

ITU-T

- Telecommunication applications
  - Speech codecs, echo cancellers, etc.

- In general
  - Speech oriented
  - Mean Opinion Score (MOS) based
  - Mostly narrowband
    - 300 – 3400 Hz
  - Wideband appearing
    - 100 – 7000 Hz
  - Naïve assessor
    - N = 12...36

- ITU-T P.800
  - Primary telecommunication listening test standard

- Covers a number of methods
  - Absolute Category Rating (ACR)
  - Comparison Category Rating (CCR)
  - Degradation Category Rating (DCR)

ITU–T P.800 (ACR)

- ✓ Absolute category rating
  - ✓ Single stimulus method
- ✓ Dependent variable
  - ✓ 5-point categorical scale
    - ✓ Listening quality
    - ✓ Listening effort
    - ✓ Loudness preference
- ✓ Independent variables
  - ✓ System/codec, speech sample, talker gender, sentence, listening level
- ✓ Naïve subjects
  - ✓ N = 24–36
- ✓ ANOVA based analysis

ITU–T P.800 (CCR)

- ✓ Comparison category rating
  - ✓ Paired comparison, hidden reference
- ✓ Dependent variable
  - ✓ 7–point categorical scale
- ✓ Independent variables
  - ✓ System/codec, speech sample, talker
- ✓ Naïve subjects
  - ✓ N = 24 – 32
- ✓ ANOVA based analysis
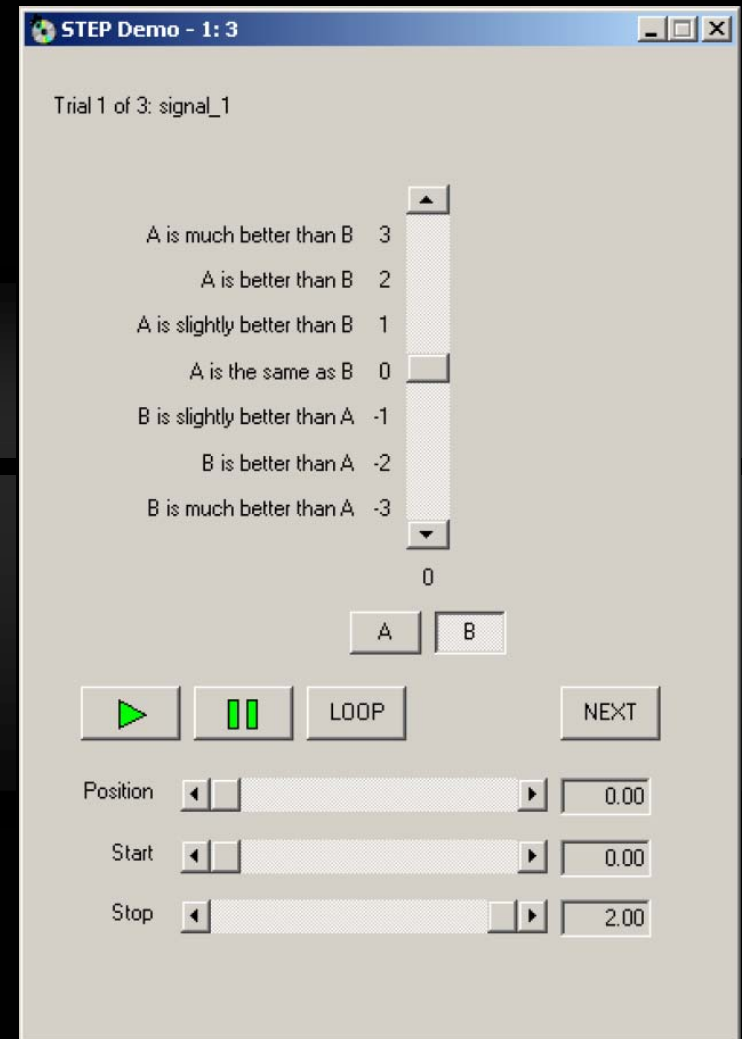
ITU-T P.800 (DCR)

- ✓ Degradation category rating
  - ✓ Fixed reference paired comparison
- ✓ Dependent variable
  - ✓ 5-point degradation categorical scale
- ✓ Independent variables
  - ✓ System/codec, speech sample, talker, background
- ✓ Naïve subjects
  - ✓ N = 32
- ✓ ANOVA based analysis

ITU-R

- ✓ Radio communication section

- ✓ Audio applications
  - ✓ E.g. audio codecs
- ✓ Full band audio applications
- ✓ Basic audio quality (BAQ) based
- ✓ Expert assessors
  - ✓ N = ~20

- ✓ A number of standards

- ✓ Two key standards
  - ✓ ITU-R BS.1116-1
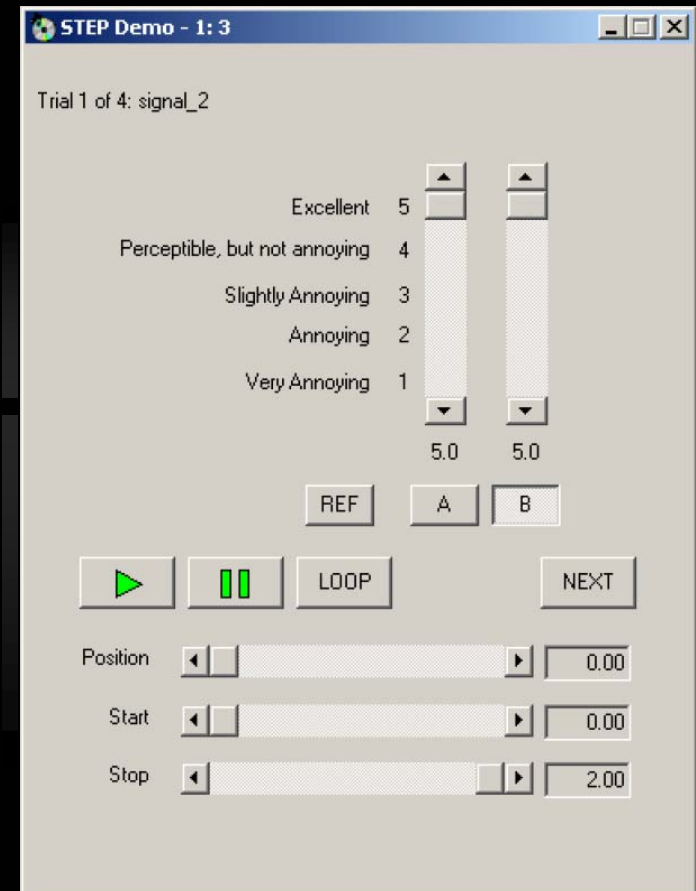
  - ✓ ITU-R BS.1534-1
   - ✓ (aka MUSHRA)

ITU–R Overview
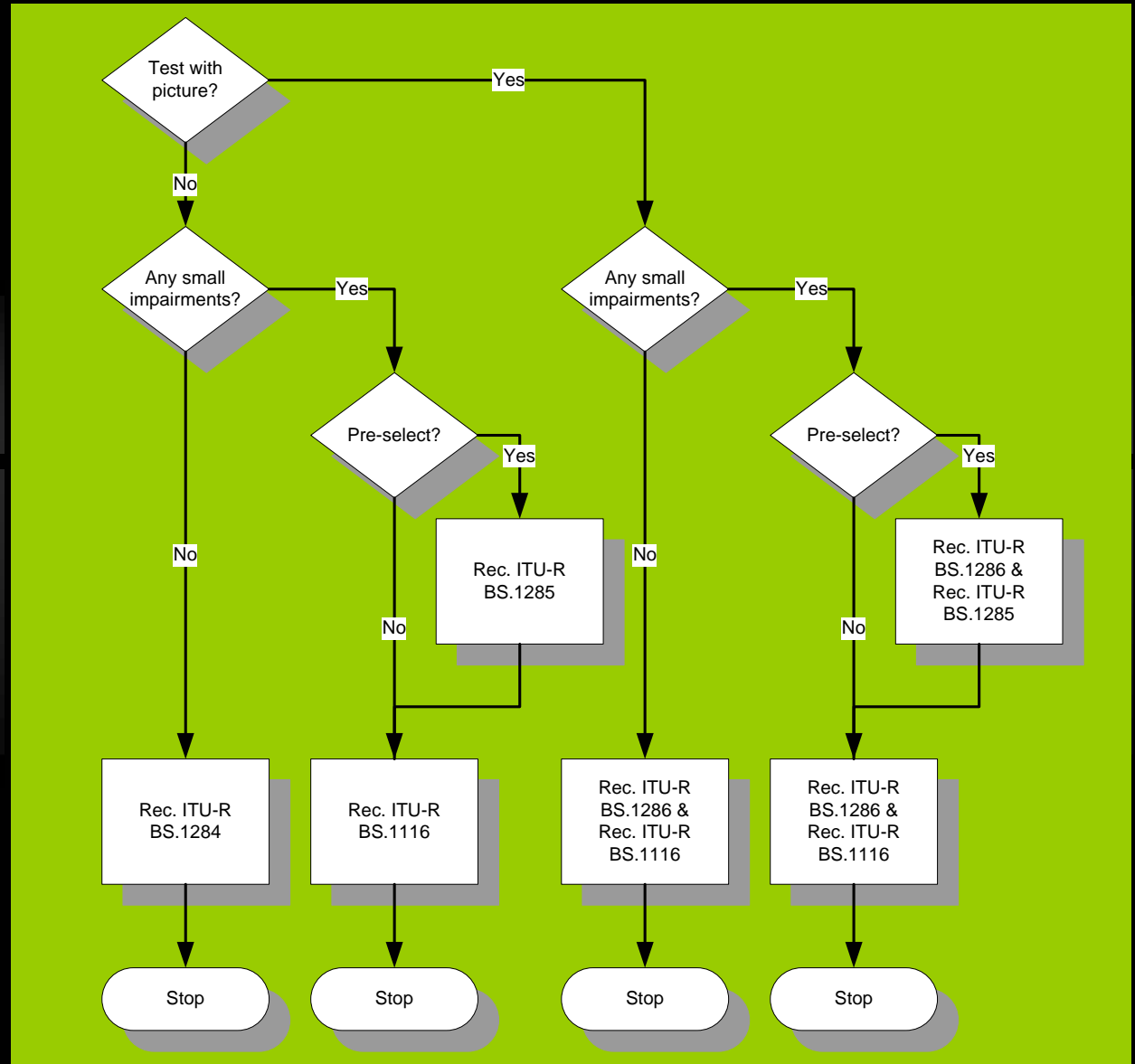
ITU-R BS.1116-1

- ✓ Evaluation of small impairment (only)
- ✓ Double blind, triple stimulus hidden reference
- ✓ Dependent variable
  - ✓ 5-point continuous rating scale
    - ✓ Basic Audio Quality (BAQ)
    - ✓ Stereophonic image quality
    - ✓ Front image quality
    - ✓ Impression of surround quality
- ✓ Independent variables
  - ✓ System/codec, programme, subject
- ✓ Expert assessors
  - ✓ Selection process is defined
  - ✓ N = 20
- ✓ ANOVA based analysis

- ✓ Listening room definition
- ✓ Loudspeaker setup definition

ITU-R BS.1534-1 (aka MUSHRA)

- ✓ Double-blind multi-stimulus with hidden reference and hidden anchors
- ✓ Dependent variable
  - ✓ 0-100 continuous quality scale (CQS) with 5 equal intervals
    - ✓ Basic Audio Quality (BAQ)
    - ✓ Stereophonic image quality
    - ✓ Front image quality
    - ✓ Impression of surround quality
- ✓ Independent variables
  - ✓ System/codec, programme, subject
- ✓ Partially screen subjects
  - ✓ N > 20
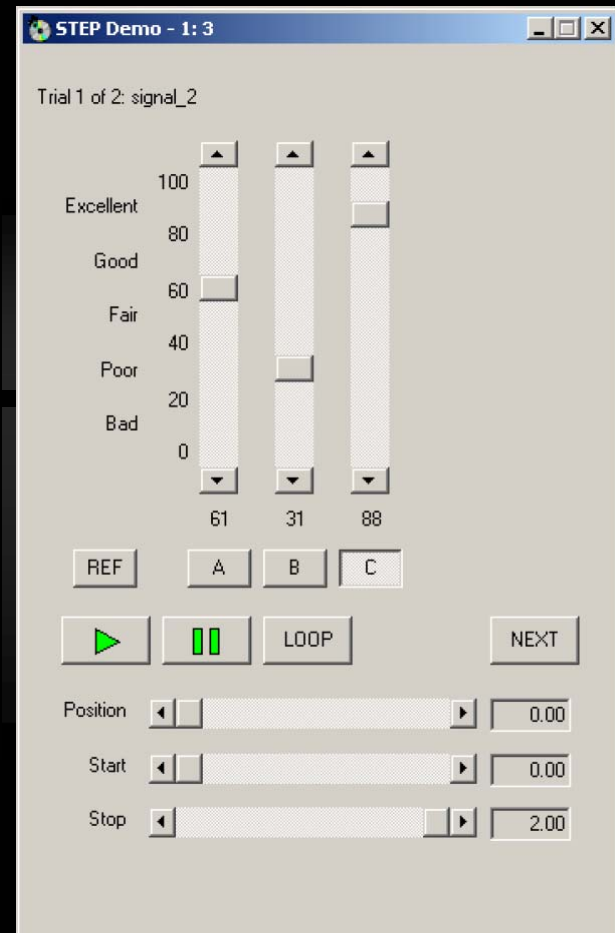
Thank you!

To learn more…

Søren Bech

Bang & Olufsen

sbe@bang-olufsen.dk

Nick Zacharov

DELTA, SenseLab

NVZ@delta.dk

Perceptual Audio Evaluation

Theory, Method and Application

Søren Bech   Nick Zacharov

WILEY