PhD Course Lecture
Subjective Assessment

Howell Istance

De Montfort University

This work is supported by the COGAIN NoE
EU IST 6th framework program

Overview

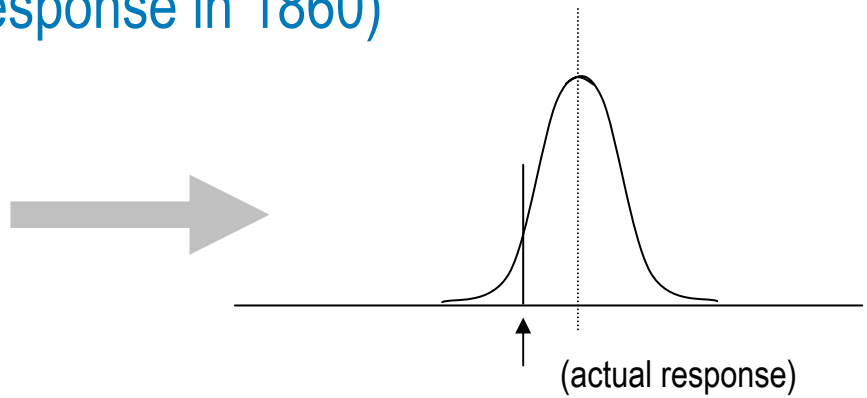
- Thurstone Scaling of subjective responses
- Rating Scales for assessing pointing device performance

Stimulus and response

- Physical stimulus invokes a subjective response
  - Luminance of a light source invokes sensation of *brightness*
  - Frequency of a sound invokes sensation of *pitch*
- A statement of opinion invokes an *attitude* towards that psychological object
- Even if the stimulus is constant, the response will vary

Thurstone's big ideas… (back in 1927)

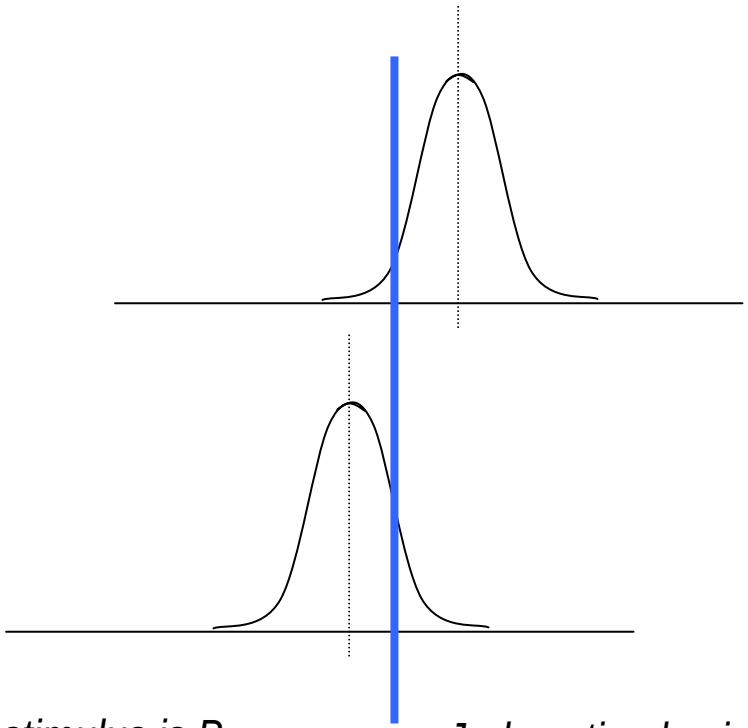
1. The response to the stimulus invoked when asked to make a judgment of some attribute lies on a psychological or subjective continuum
- The probability of a particular level of response is described by the normal distribution
- (not really Thurstone's own idea, Fechner had mentioned variability in response in 1860)



stimulus

response continuum

Judgement threshold



Response to stimulus A

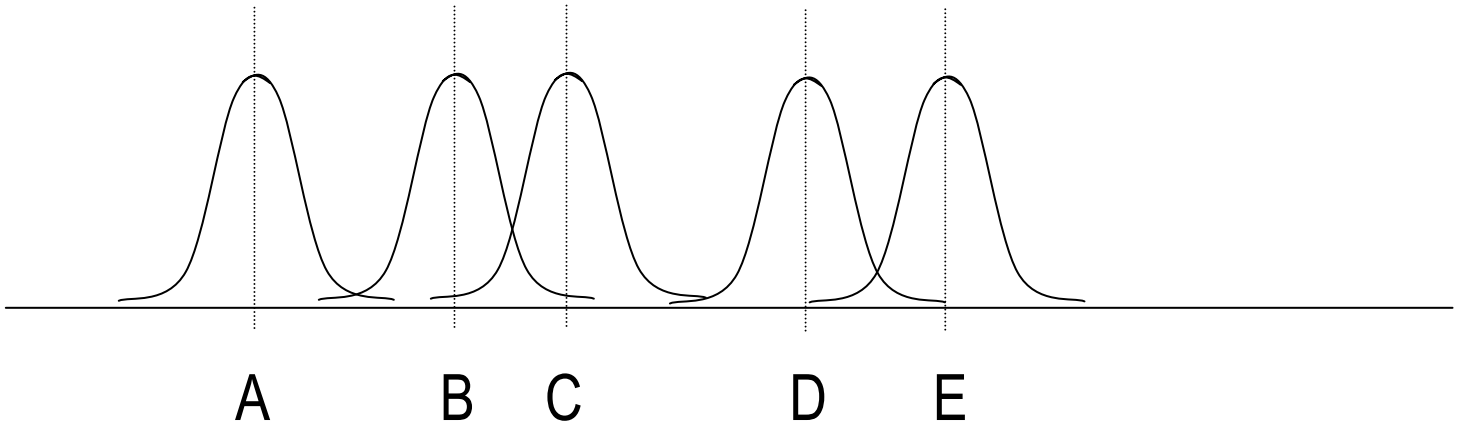
Response to stimulus B

Judge stimulus is B

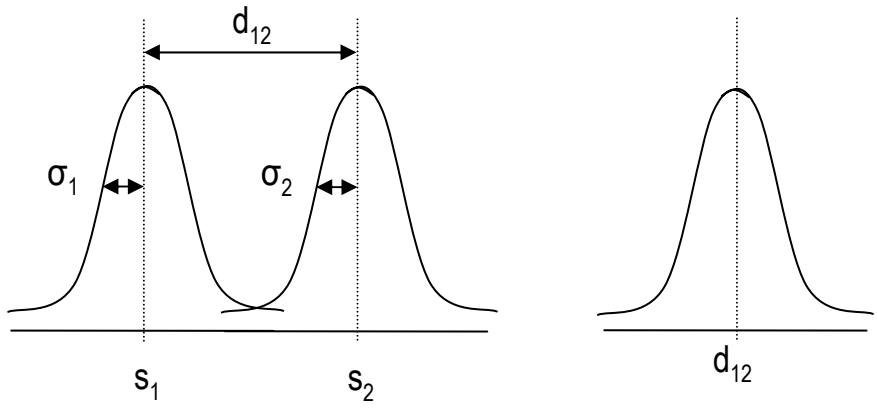
Judge stimulus is A

So what…?

- We often want to be able to derive a interval scale where the positions of the modal values of the responses to different stimuli are known



Consider distribution of differences between items drawn from 2 populations



Mean	S_1	S_2	$d_{12} = S_1 - S_2$
Variance	σ_1^2	σ_2^2	$\sigma_1^2 + \sigma_2^2 + 2\rho\,\sigma_1\,\sigma_2$ (variance sum law)

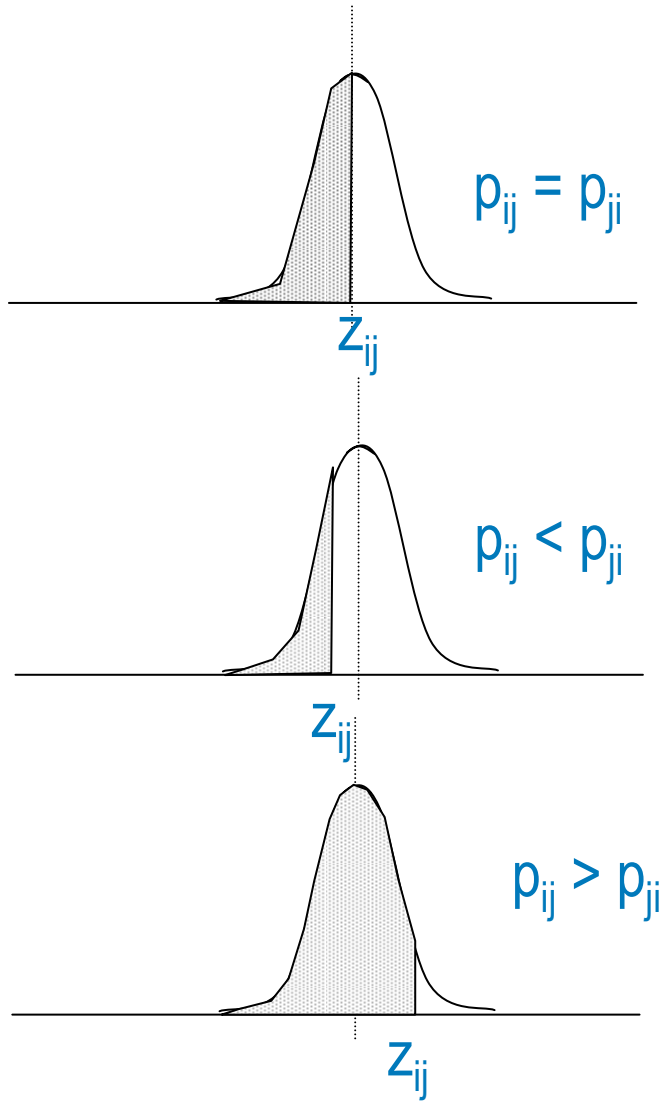
Thurston's Law of Comparative Judgment

- Suppose i and j are two statements about some psychological object
- Ask a large number of people to judge whether i or j is more favourable towards the object
- A proportion of these will say i is more than j (p_{ij}) and a proportion will say j is more than I (p_{ji})
- If $p_{ij} = p_{ji} = 0.5$ then we can assume that the modal points of the two distributions overlap and the scale points are the same

Scale separation between modal responses to i and j

$s_i - s_j = z_{ij} / \mathrm{sqrt}(\sigma_i^2 + \sigma_j^2 + 2\rho\,\sigma_i\,\sigma_j)$

where the normal deviate z_{ij} corresponds to the proportion of comparative judgments p_{ij}



Some assumptions about the distribution..

$s_i - s_j = z_{ij} / \mathrm{sqrt}(\sigma_i^2 + \sigma_j^2 + 2\rho\,\sigma_i\,\sigma_j)$

- Lets assume the distributions have the same variance..
- $\sigma_i^2 + \sigma_j^2 = \sigma^2$

- Lets assume that the responses to each stimulus are independent (i.e. a high response to S_i is not associated with a high response to S_j) … so $\rho = 0$

$s_i - s_j = z_{ij} / \mathrm{sqrt}(\sigma_i^2 + \sigma_j^2 + 2\rho\,\sigma_i\,\sigma_j)$ ➡ $s_i - s_j = z_{ij} / 1.414\,\sigma$

Deriving a scale of the taste of (British) beer

- Several ways of using the law of comparative judgments to obtain an interval scale of on which responses to stimuli can be placed
- Such as the method of paired comparisons.
- Give lots of people all pairs of stimuli, and ask for judgement, which is better (taller, nicer, is more favourable..)
- In the case of beer, “which beer tastes better..? “

Single subject response.. (S1 better than S2 = 1)

S2 \ S1	Marstons Pedigree	Courage Directors	Brains SA	Bass	Ruddles County	Sam Smiths	Greene King	Theakston's Bitter
Marstons Pedigree	x	0	0	0	0	0	0	0
Courage Directors	1	x	0	1	0	0	0	1
Brains SA	1	0	x	1	0	0	0	1
Bass	0	0	0	x	0	0	0	0
Ruddles County	1	1	1	1	x	1	1	1
Sam Smiths	1	0	1	1	0	x	0	1
Greene King	1	1	1	1	1	1	x	1
Theakston's Bitter	1	0	0	1	0	0	0	x

Proportions preferring S1 to S2

S2 \ S1	Marstons Pedigree	Courage Directors	Brains SA	Bass	Ruddles County	Sam Smiths	Greene King	Theakston's Bitter
Marstons Pedigree	x	.10	.05	.20	.05	.02	.01	.01
Courage Directors	.90	x	.20	.98	.05	.06	.02	.60
Brains SA	.95	.80	x	.80	.10	.10	.05	.95
Bass	.80	.02	.20	x	.20	.10	.10	.20
Ruddles County	.95	.95	.90	.80	x	.95	.05	.95
Sam Smiths	.98	.94	.90	.90	.05	x	.20	.98
Greene King	.99	.98	.95	.90	.95	.98	x	.95
Theakston's Bitter	.90	.40	.05	.80	.05	.02	.05	x

Convert Proportions to z-scores

S2 \ S1	Marstons Pedigree	Courage Directors	Brains SA	Bass	Ruddles County	Sam Smiths	Greene King	Theakston's Bitter
Marstons Pedigree	x	-1.28	-1.64	-0.85	-1.64	-2.05	-2.33	-1.28
Courage Directors	1.28	x	-0.85	2.05	-1.64	-1.55	-2.05	0.25
Brains SA	1.64	0.85	x	0.85	-1.28	-1.28	-1.64	1.64
Bass	0.85	-2.05	-0.85	x	-0.85	-1.28	-1.28	-0.85
Ruddles County	1.64	1.64	1.28	0.85	x	1.64	-1.64	1.64
Sam Smiths	2.05	1.55	1.28	1.28	-1.64	x	-2.05	2.05
Greene King	2.33	2.05	1.64	1.28	1.64	2.05	x	1.64
Theakston's Bitter	1.28	-0.25	-1.64	0.85	-1.64	-2.05	-1.64	x

Average z-scores to give interval scale

S2 \ S1	Marstons Pedigree	Courage Directors	Brains SA	Bass	Ruddles County	Sam Smiths	Greene King	Theakston's Bitter
Marstons Pedigree	x	-1.28	-1.64	-0.85	-1.64	-2.05	-2.33	-1.28
Courage Directors	1.28	x	-0.85	2.05	-1.64	-1.55	-2.05	0.25
Brains SA	1.64	0.85	x	0.85	-1.28	-1.28	-1.64	1.64
Bass	0.85	-2.05	-0.85	x	-0.85	-1.28	-1.28	-0.85
Ruddles County	1.64	1.64	1.28	0.85	x	1.64	-1.64	1.64
Sam Smiths	2.05	1.55	1.28	1.28	-1.64	x	-2.05	2.05
Greene King	2.33	2.05	1.64	1.28	1.64	2.05	x	1.64
Theakston's Bitter	1.28	-0.25	-1.64	0.85	-1.64	-2.05	-1.64	x
Σ z	11.07	2.51	-.078	6.31	-7.05	-4.52	-12.63	5.09
mean z	1.38	0.31	-0.1	0.79	-0.88	-0.56	-1.58	0.63
mean z + 1.58	2.96	1.89	1.48	2.37	0.7	1.02	0	2.21

Ummm!
☺

Uuugh!
☹

Other approaches..

- Paired comparisons quickly becomes impractical when the number of items to be scaled increases..
- We want to often find the scale points of many statements or objects
- Method of Equally Appearing Intervals requires only one judgment per statement by 1 subject.

Equally Appearing Intervals

frequency	2	2	6	2	6	12	44	26	68	28	4
scale value	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*

extremely unfavourable neutral extremely favourable

- "The British Army has a hard job to do in Basra"
- Ask 200 people to rate how favourable or otherwise the statement is towards British foreign policy in Iraq

  …regardless of their own agreement with the statement

Calculate a scale value…

- Based on the frequency distribution and the median value, we can derive a scale value and interquartile range for this statement on the continuum of 'favourableness'

- Scale value $= l + ((0.5 - \Sigma p_b)/p_w))$

- l = lower limit of interval containing median
- p_b = proportion below interval
- p_w = proportion within interval

Attitude scale construction

- We often need to measure attitudes towards a particular psychological object
- Generate lots of statements about the object
- Find the position of each statement on a scale of ‘favourableness’ towards the object.
- Select a subset of statements with low variability and scale positions dispersed along the continuum
- Use this subset to conduct an attitude survey amongst groups of interest

Rating scales for pointing devices

- Rating scale = quantisation of a psychological continuum
- Attributes
  - Fully labelled/part labelled
  - Number of points
  - Description of quantifiers
  - Unipolar (none – very X) vs Bipolar (very unX – very X)

Examples of unipolar and bipolar scales

- How fast was the pointing device?
- *(Bipolar) Extremely Slow – Extremely fast*
- *(Unipolar) Not at all fast – Extremely fast*

Very clumsy English!

- How tired do you feel ?
- *(Unipolar) Not at all tired – Extremely tired*
- *(Bipolar) Extremely not tired – Extremely tired*

What scale is best?

- Is the NASA-TLX the best way to rate aspects of workload and performance?
- The scale used needs to be validated for use with pointing devices
- Eyetrackers usually introduce noise and inaccuracy to the location of a screen pointer.
- Performance tests are usually of short duration (so the rating scale needs to be sensitive to this)

Why is this important?

- Comparisons between 2 devices may result in ‘no significant difference’
- This may be because there isn’t one…
- Or, because the measuring device used isn’t sensitive enough to detect the difference.

- We need to be confident that a result of ‘no difference’ is due to the first of these cases.

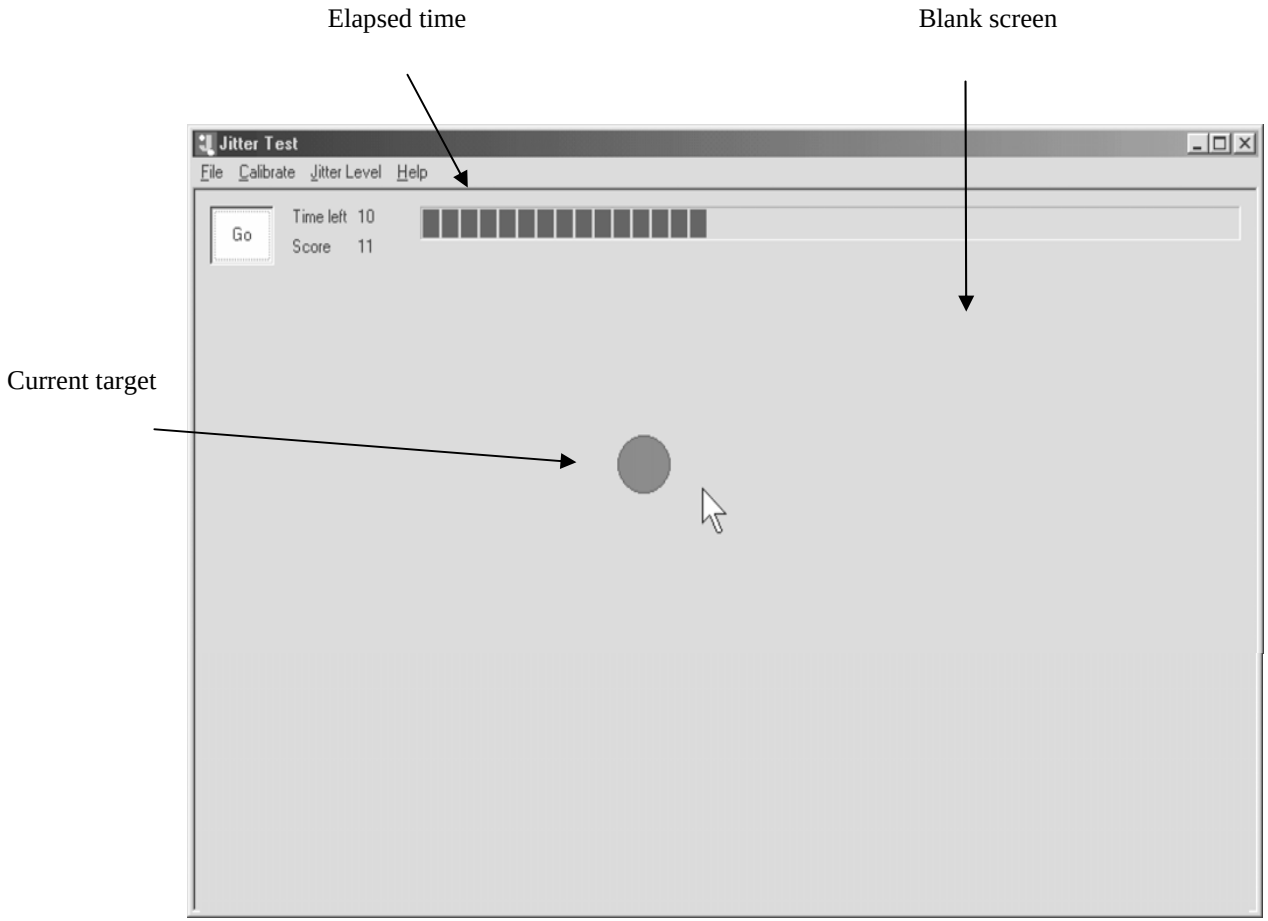
Jitter test

- Target acquisition test
- Constant Fitts Index of Difficulty of 3.2

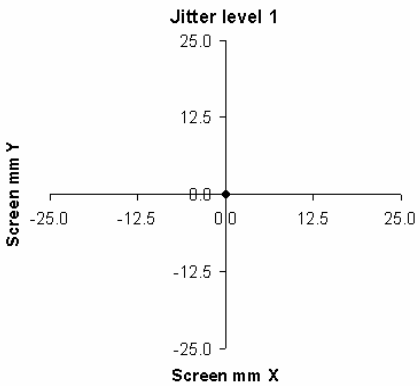
- New target appears on acquisition of previous one, random direction, constant distance and same size

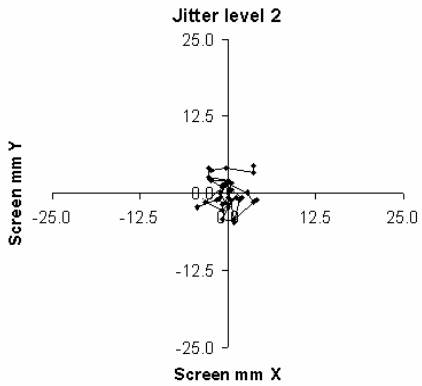
- Jitter introduced into pointer response to mouse movement

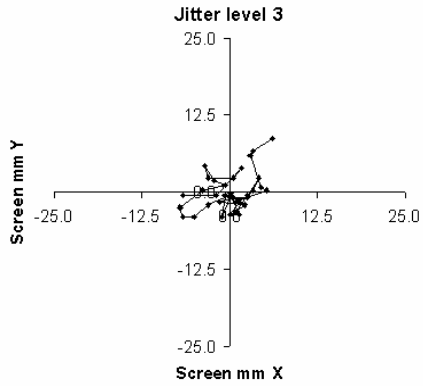
Jitter test screen shot

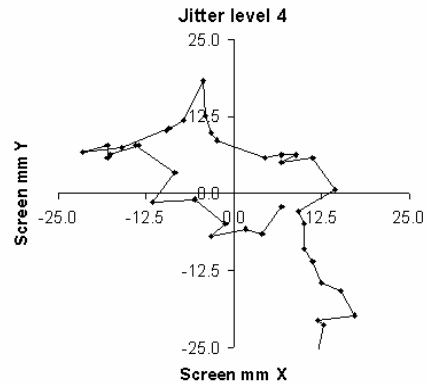


Jitter conditions









1 (none)

2 (smallest perceivable difference)

3

4 (lots)

Experiment to compare scales

- 17 candidate scales compared
  - 5 point /fully labelled/partially labelled bipolar/unipolar
  - 7 point /fully labelled/partially labelled bipolar/unipolar
  - 9 point /fully labelled/partially labelled bipolar/unipolar
  - 11 point /fully labelled/partially labelled bipolar/unipolar

  - 20 interval partially labelled unipolar NASA tlx

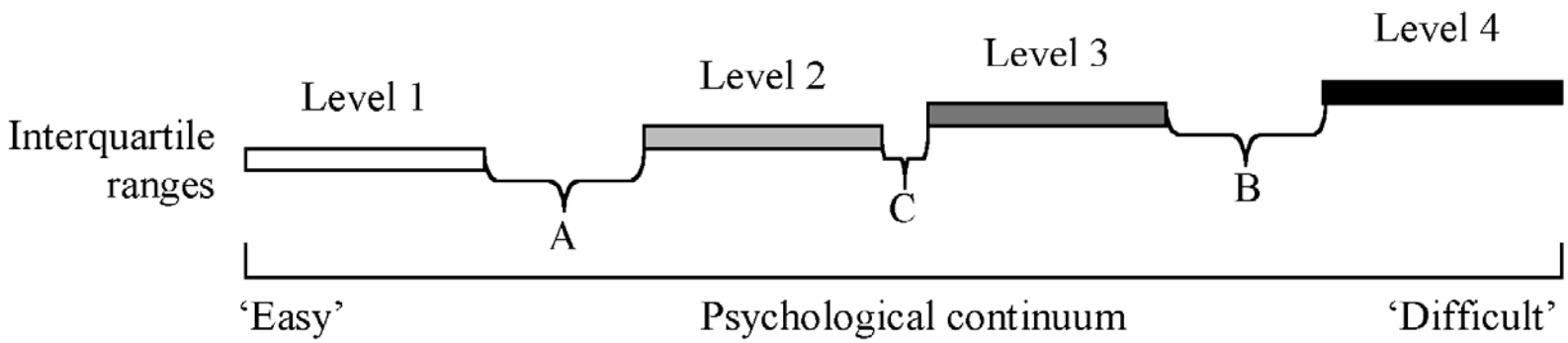
5 point fully labelled bipolar scale

Extremely easy	Easy	Neither easy nor difficult	Difficult	Extremely difficult

Experiment…

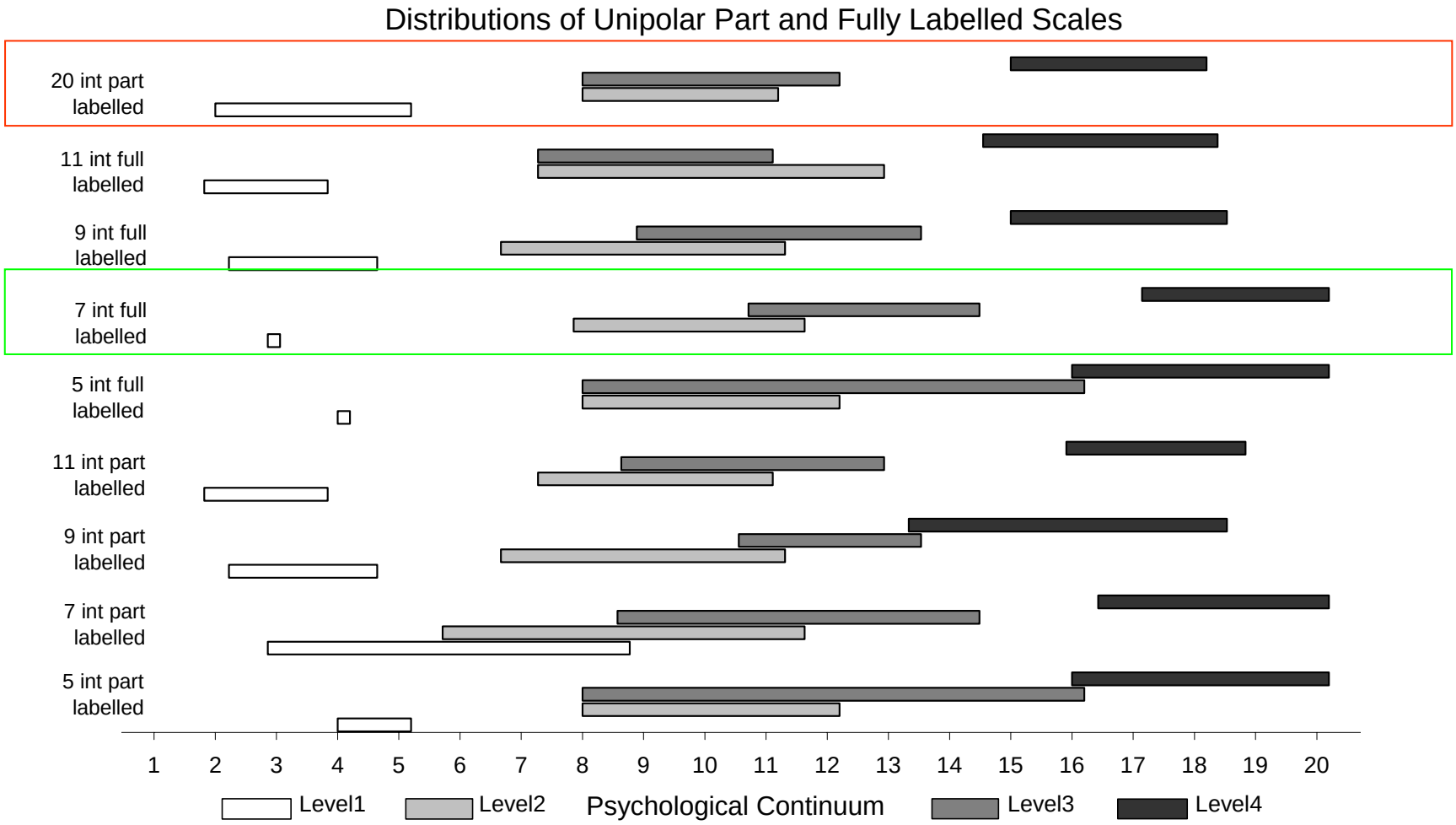
- 40 subjects (35 m, 5 f)
- Each condition, 30 second chase-the-target jitter test
- 4 levels of jitter
- 17 scales
- Ask the subject to rate how easy the task was, for each level of jitter on each scale

Ideal result

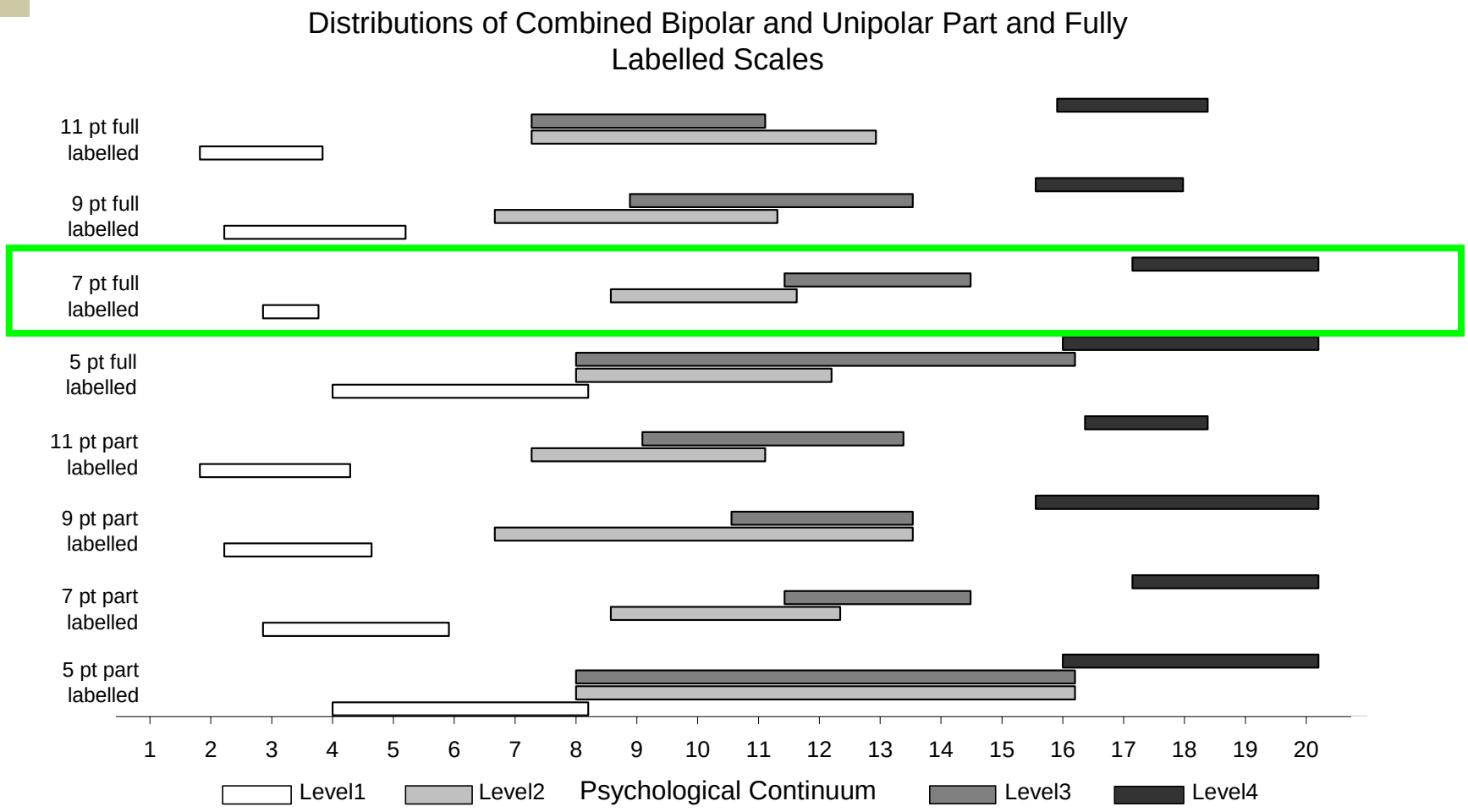


- Levels ordered correctly
- Level 2 and level 3 separate - no overlap
- C much smaller than A and B

Actual results – unipolar scales



Combined unipolar and bipolar results



Example of using the scale..

Workload Assessment Questionnaire

Please circle the 'X' closest to your opinion
← low workload ratings high →

1. How much *physical* effort or activity was required to operate the system?

Extremely low physical effort	Considerably low physical effort	Somewhat low physical effort	Neither high nor low physical effort	Somewhat high physical effort	Considerably high physical effort	Extremely high physical effort

2. How much *mental* effort or concentration was required to operate the system?

Extremely low mental effort	Considerably low mental effort	Somewhat low mentall effort	Neither high nor low mental effort	Somewhat high mental effort	Considerably high mental effort	Extremely high mental effort

3. How much…..