Psychophysics 101

How to run perception experiments in computer graphics

SIGGRAPH 2008 Courses Program

Course organizer and presenter
James A. Ferwerda
Munsell Color Science Laboratory
Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology

Course description

Psychophysical methods from experimental psychology can be used to quantify the relationships between the properties of images and what people perceive. The results of psychophysical experiments can be used to create predictive models of human perception that can guide the development of effective and efficient graphics algorithms and enabling graphical interfaces. This course will provide an introduction to the use of psychophysical methods in computer graphics and will teach attendees how to develop experiments that can be used to advance graphics research and applications. Throughout the presentation, graphics-relevant examples will be used so attendees will understand how to design and run their own experiments; analyze the results; and develop perceptually-based algorithms and applications. This course will be of interest to members of the graphics community who want to be able to interpret the results of perception psychology experiments and develop their own user studies of computer graphics techniques.

Prerequisites

This course assumes a basic level understanding of issues in computer graphics and electronic imaging. Familiarity with freshman-level college math will be helpful. *No specific knowledge of perception psychology or statistical methods will be required. All relevant concepts will be introduced in the class.*

Syllabus

- **Welcome, Introductions, Schedule Review (5 mins)**
    - Motivation / Orientation (20 mins)
    - Psychophysical Methods (50 mins)
    - Case Studies (25 mins)
    - Summary/Resources(5 mins)

- **Motivation/Orientation (20 mins)**
    - Why does graphics need psychophysics?
    - What kinds of problems can be addressed by psyshophysics?
    - How does psychophysics relate to physical measurement?
    - How is psychophysics different from usability?
    - What kinds of results can psychophysics produce?
    - Why don’t we just mine the existing literature?
    - How do we make progress?
    -
- **Psychophysical Methods (50 mins)**
    - **Introduction**
        - need for objective metrics of subjective visual experience
        - fundamental psychophysical metrics: **thresholds** and **scales**
        - examples of threshold and scale metrics in computer graphics
    - **Methods for measuring <u>thresholds</u>**
        - the method of adjustment
        - the method of limits
        - the method of constant stimuli

  - **Threshold models**
    - psychometric functions
  - **Sources of error in threshold measurements**
  - **Signal detection theory**
    - SR matrices, ROC curves
    - error-free threshold measures
  - **Suprathreshold scaling methods**
    - Types of psychophysical scales
      - nominal, ordinal, interval, ratio
    - Ordinal and interval scaling methods
    - Thurstonian scaling
    - Ratio scaling methods
  - **Scaling models**
    - Weber's law
    - Steven's power law
  - **Multidimensional scaling (MDS)**
    - theory, demo, data analysis
  - **Practicalities of running psychophysical experiments in computer graphics**
    - selecting test stimuli
    - selecting observers
    - the importance of observer instructions
    - presentation/user interface issues
    - collecting data

- **Case Studies of Psychophysical Methods in Graphics (25 mins)**
  - Brief descriptions and critique of past and current research in perceptually-based graphics. Topic areas:
    - realistic image synthesis
    - geometric simplification
    - image-based modeling
    - compression
    - virtual environments
    - visualization

- **Summary/Conclusions (5 mins)**
    - utility and limits of psychophysical methods
    - relevance for computer graphics
  - Resources
    - books, papers, journals, web sites
    - software packages
    - organizations, conferences

Resources

Books

Gescheider, G.A. (1997) Psychophysics: The Fundamentals, 3rd Edition. Erlbaum.

Bartelson, C.J. and Grum, F. (Eds.) (1984) Optical Radiation Measurements, Vol 5: Visual Measurements. Academic Press, New York.

Engeldrum, P.G. (2000) Psychometric scaling: A Toolkit for Imaging Systems Development. Imcotek Press.

Guilford, J.P. (1954) Psychometric methods. Mcgraw-Hill.

Torgerson, W.S. (1960) Theory and Methods of Scaling. Wiley.

Green, D.M. and Swets, J.A. (1966) Signal Detection Theory and Psychophysics. Wiley.

Fechner, G.T. (1966) Elements of Psychophysics. Holt, Rinehart & Winston

Papers/Standards

ASTM (American Society for Testing and Materials), Standard Guide for Conducting Visual Experiments, E1808-96

ASTM (American Society for Testing and Materials), Standard Guide for Selection, Evaluation, and Training of Observers, E1499-94

CIE Technical Committee 1-34 Testing Color-Appearance Models: Guidelines for Coordinated Research - Alessi, P.J. (1994) Color Research and Applications,19, 48-58.

Use of computers and cathode-ray-tube displays in visual psychophysics - special issues of the journal Spatial Vision 10(4) and 11(1) - http://www.hans.strasburger.de/

Software/Websites

Psychophysics Toolbox (Matlab-based)
http://psychtoolbox.org/

Psychophysica/Cinematica (Mathematica/Quicktime-based)
http://vision.arc.nasa.gov/mathematica/psychophysica/

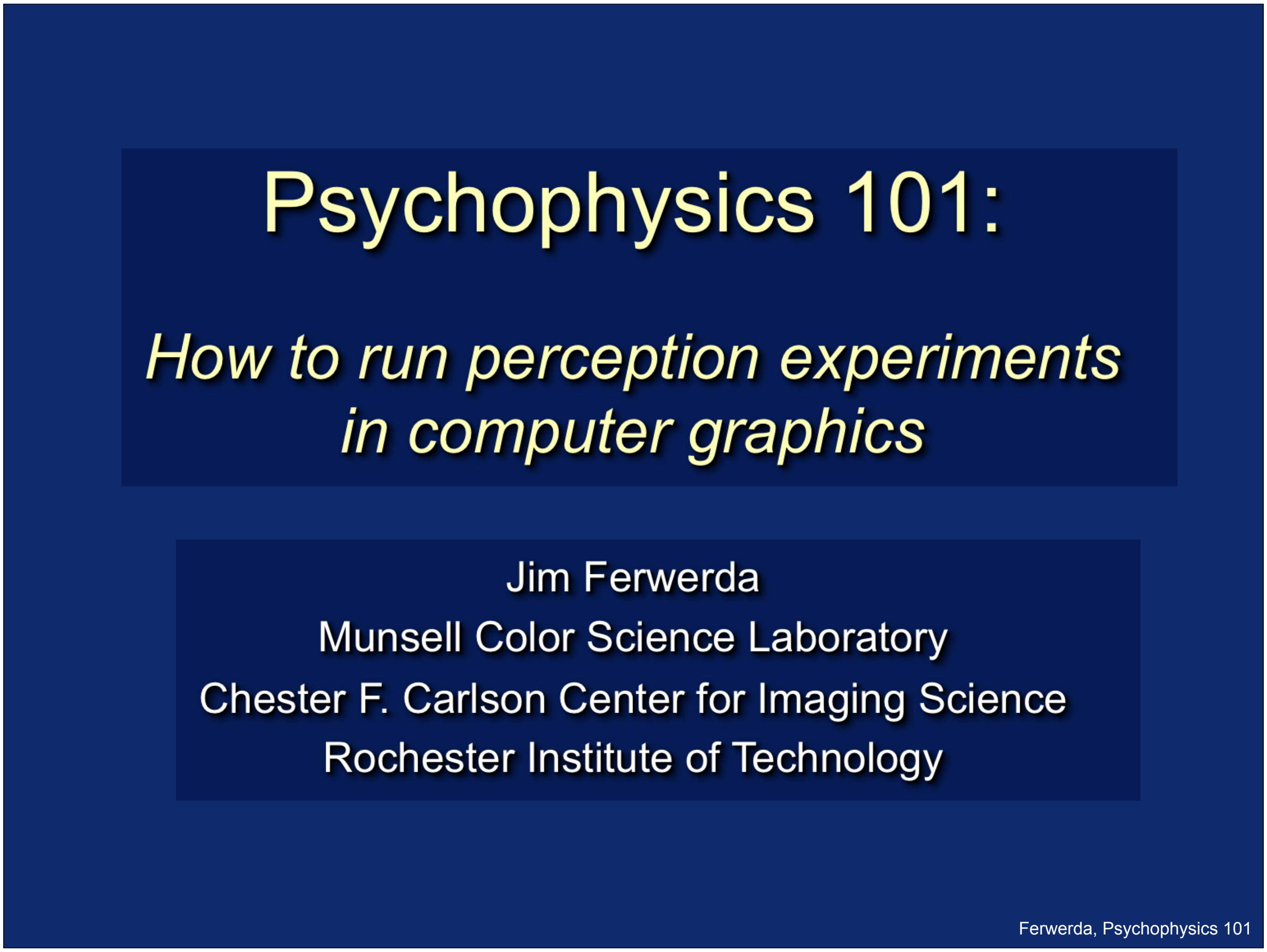
Strasburger's overview of psychophysics software
http://www.hans.strasburger.de/

Course presenter biographical sketch

James A. Ferwerda is an Associate Professor in the Munsell Color Science Laboratory in the Center for Imaging Science at the Rochester Institute of Technology. He received a B.A. in Psychology, M.S. in Computer Graphics, and a Ph.D. in Experimental Psychology, all from Cornell University. The focus of his research is on building computational models of human vision from psychophysical experiments, and developing advanced graphics algorithms based on these models. Current research interests include: high dynamic range imaging; perceptually-based rendering; perception of material properties; and low vision and assistive technologies. Prior to joining the Faculty at RIT he was a Research Associate in the Program of Computer Graphics at Cornell. In 1992 he received the IEEE Computer Society Paper of the Year Award, and in 2003 he was selected for the National Academy of Engineering Frontiers of Engineering Program. He is an Associate Editor of ACM Transactions on Applied Perception, was Guest Editor for a special edition of IEEE Computer Graphics and Applications on Applied Perception, and serves as a member of CIE Technical Committee TC8-08 on High Dynamic Range Imaging.

Contact information

James A. Ferwerda, Ph.D.
Associate Professor
Munsell Color Science Laboratory
Chester F. Carlson Center for Imaging Science
18/1069 Lomb Memorial Drive
Rochester Institute of Technology
Rochester, NY 14623
585-475-4923
jaf@cis.rit.edu
http://www.cis.rit.edu/jaf

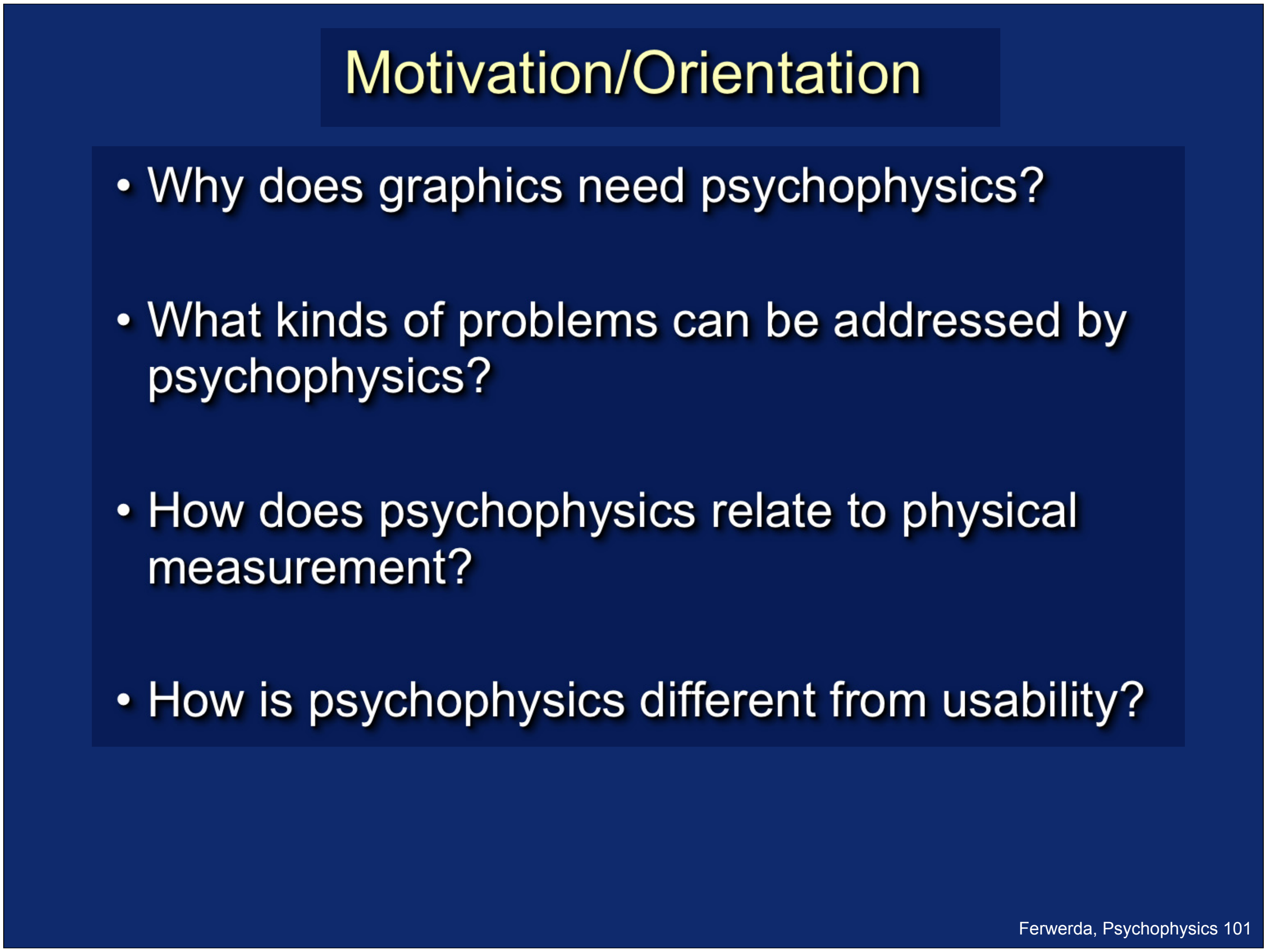


Psychophysics 101:

How to run perception experiments in computer graphics

Jim Ferwerda
Munsell Color Science Laboratory
Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology

- One of the hallmarks of a science is the development of reliable techniques for measuring the phenomena of interest. For example, in physics there are standard techniques for measuring an object’s mass or velocity. In chemistry there are techniques for measuring the strength of bonds or the energy given off by a reaction. In computer graphics, we would like to be able to measure what people perceive when they look at our images, so we can create more effective and compelling images, and create them more efficiently. However, while physical quantities can be measured with standard methods, our visual experiences are subjective and cannot be measured directly.
- If computer graphics is to become a “science of visualization” then researchers need to be able to conduct experiments that quantify the relationships between the parameters of our algorithms, the images they produce, and the visual experiences the images create. Psychophysics provides a sound set of methodologies for conducting and analyzing these experiments. This tutorial provides a survey of the issues and methodologies of visual psychophysics.



Motivation/Orientation

- Why does graphics need psychophysics?
- What kinds of problems can be addressed by psychophysics?
- How does psychophysics relate to physical measurement?
- How is psychophysics different from usability?

- graphics are generated to be useful to people
- we need to be able to determine what factors contribute to visual experience
- we need to be able to assess what methods produce an effective visual experience
- realistic image synthesis
  - how accurate does the input need to be?, what input is needed?, how accurate does the light transfer need to be?, how are the results in physical units transformed to displays?
- data visualization
  - how should data values be mapped to visual attributes?, how effective are different visual cues for conveying information about data, what are the interactions between these different cues?, how can we make sure that the images we create are faithful representations
- virtual environments
  - what trade-offs are acceptable to maintain real time performance?, what spatial representations are adequate?, what are the perceptual differences between screen-based and immersive displays?
- compression
  - what kinds of artifacts are visually acceptable in still images? In temporal sequences? In 3D geometric models?
- human observers are responding to physical stimulus
- depending on problem various physical measurements also needed
- object shape/material properties; light energy from real scenes/displays
- focus is on visual perception and computing images , but overlap with these areas

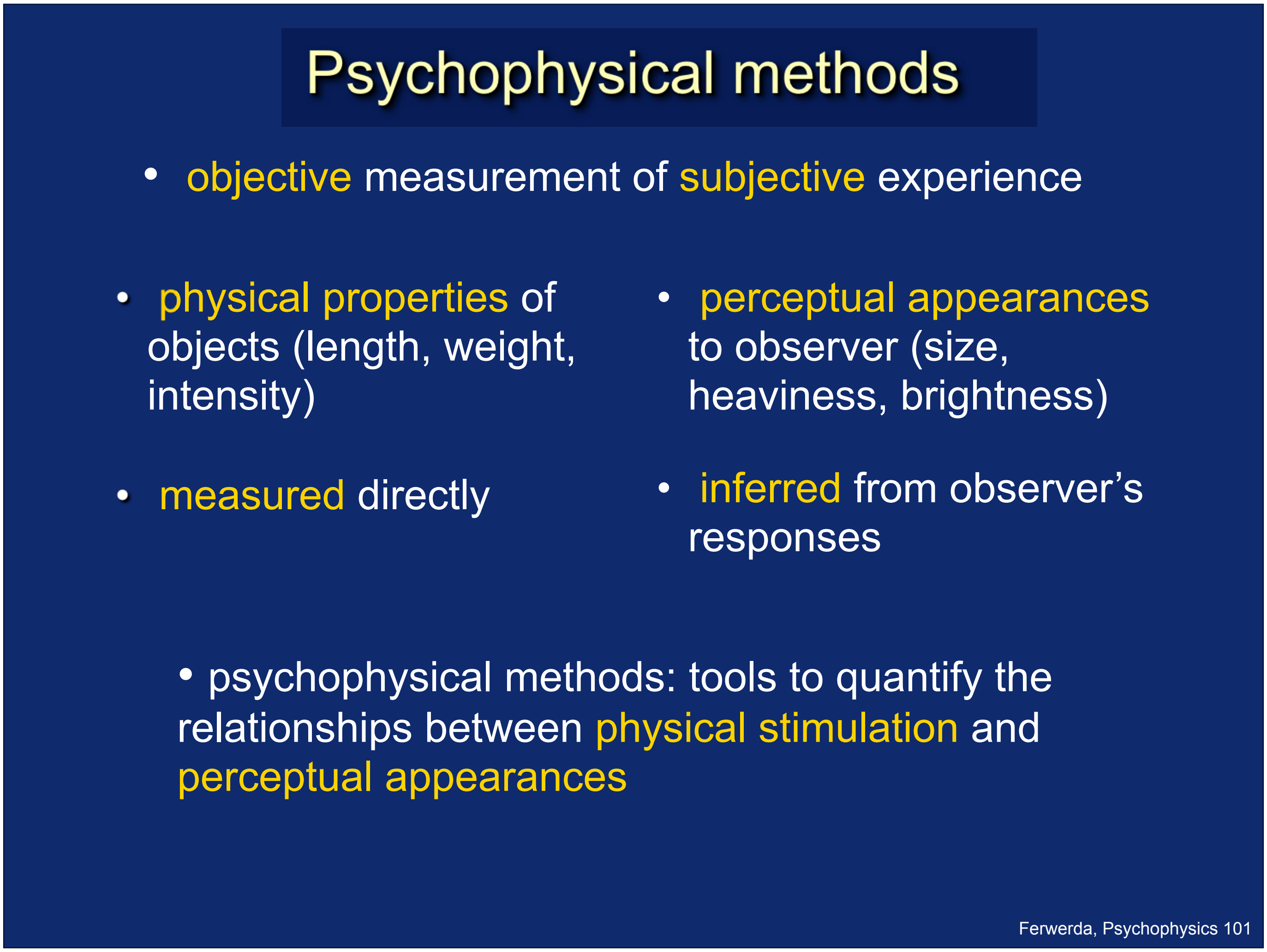
Motivation/Orientation

- What kinds of results can psychophysics produce?
- Why don’t we just mine the existing literature?
- How do we make progress?

- more efficient graphics algorithms -- computing only what is necessary
- more effective graphics methods -- choosing the right image to generate

- graphics builds on psychophysical research (e.g. colorimetry)
- goals of psychophysical research are different than graphics research
- determining contrast sensitivity vs. designing a rendering method that uses a model of contrast sensitivity

- adopt established experimental methods
- build a literature of results relevant to graphics techniques



Psychophysical methods

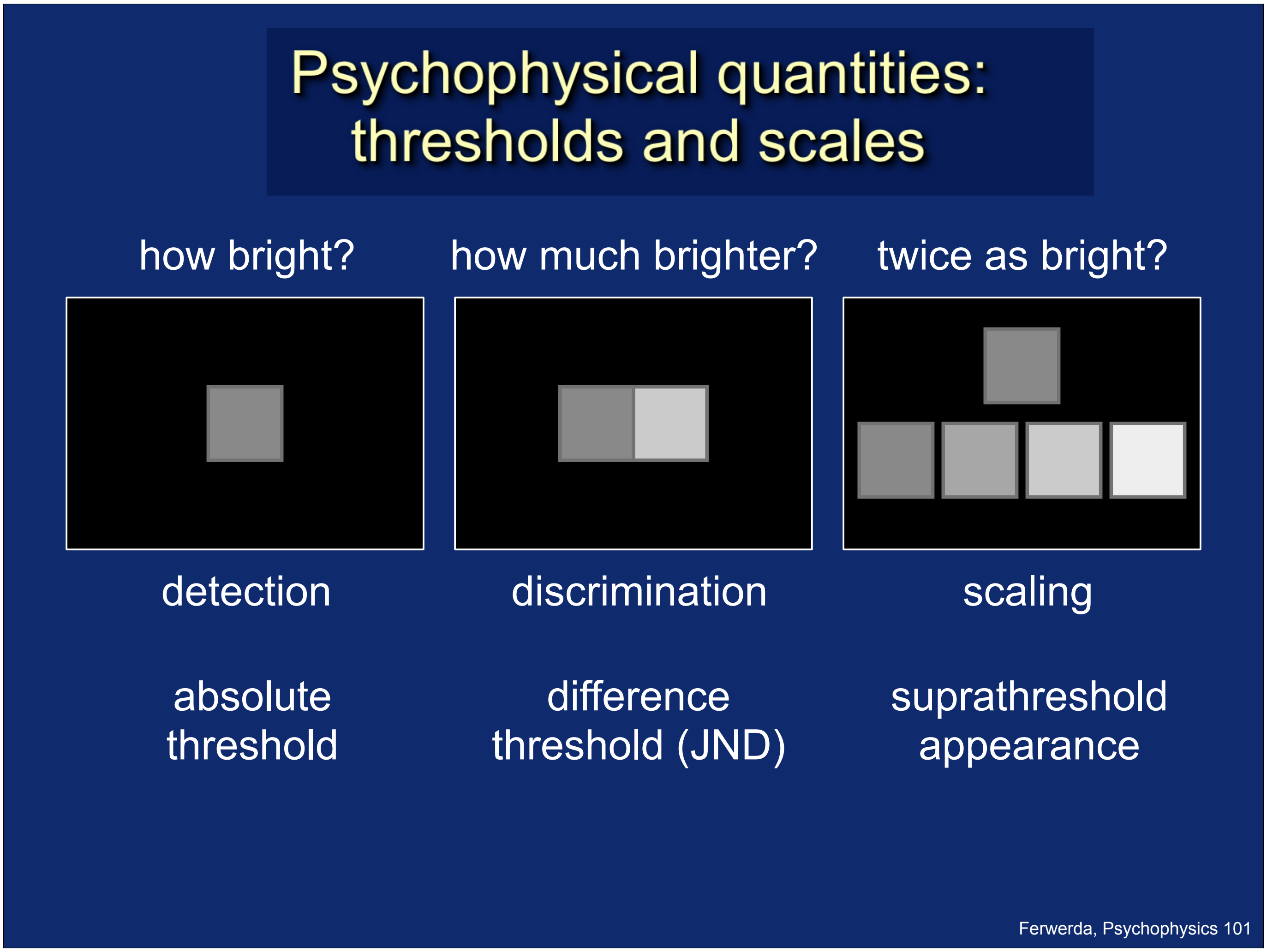
- objective measurement of subjective experience

- physical properties of objects (length, weight, intensity)
- measured directly

- perceptual appearances to observer (size, heaviness, brightness)
- inferred from observer's responses

- psychophysical methods: tools to quantify the relationships between physical stimulation and perceptual appearances

- what is psychophysics?
- branch of experimental psychology
- concerned with developing experimental methods for objectively measuring our subjective perceptual experiences
- why is this a problem?
- objects in the world have physical properties (length, weight, intensity)
- can be measured directly with instruments
- these objects create distinct perceptual impressions (size, heaviness, brightness)
- but can't put electrodes in peoples brains to directly measure sensations, so we have to infer what people perceive from their responses
- psychophysics provide a set of tools that allow us to make objective measurements of perception
- quantifies the relationships between the dimensions of the external world of physical stimulation and our internal world of perceptual appearances



Psychophysical quantities: thresholds and scales

how bright?	how much brighter?	twice as bright?
detection	discrimination	scaling
absolute threshold	difference threshold (JND)	suprathreshold appearance

- two fundamental quantities that psychophysical methods let us measure: thresholds and scales
- thresholds describe the limits of perception
- two kinds: absolute/detection threshold, difference/discrimination threshold
- detection thresholds measure the absolute limits of perception

    for example: might ask how bright does a patch need to be to be seen, photometric efficiency of vision

- difference threshold measures ability to discriminate between similar physical stimuli

    for example: how much brighter does the patch on the right need to be to be seen as different than the square on the left

- this value is known as the just-noticeable-difference (JND)
- JNDs are the basic units that form the foundation of many engineering applications of perception

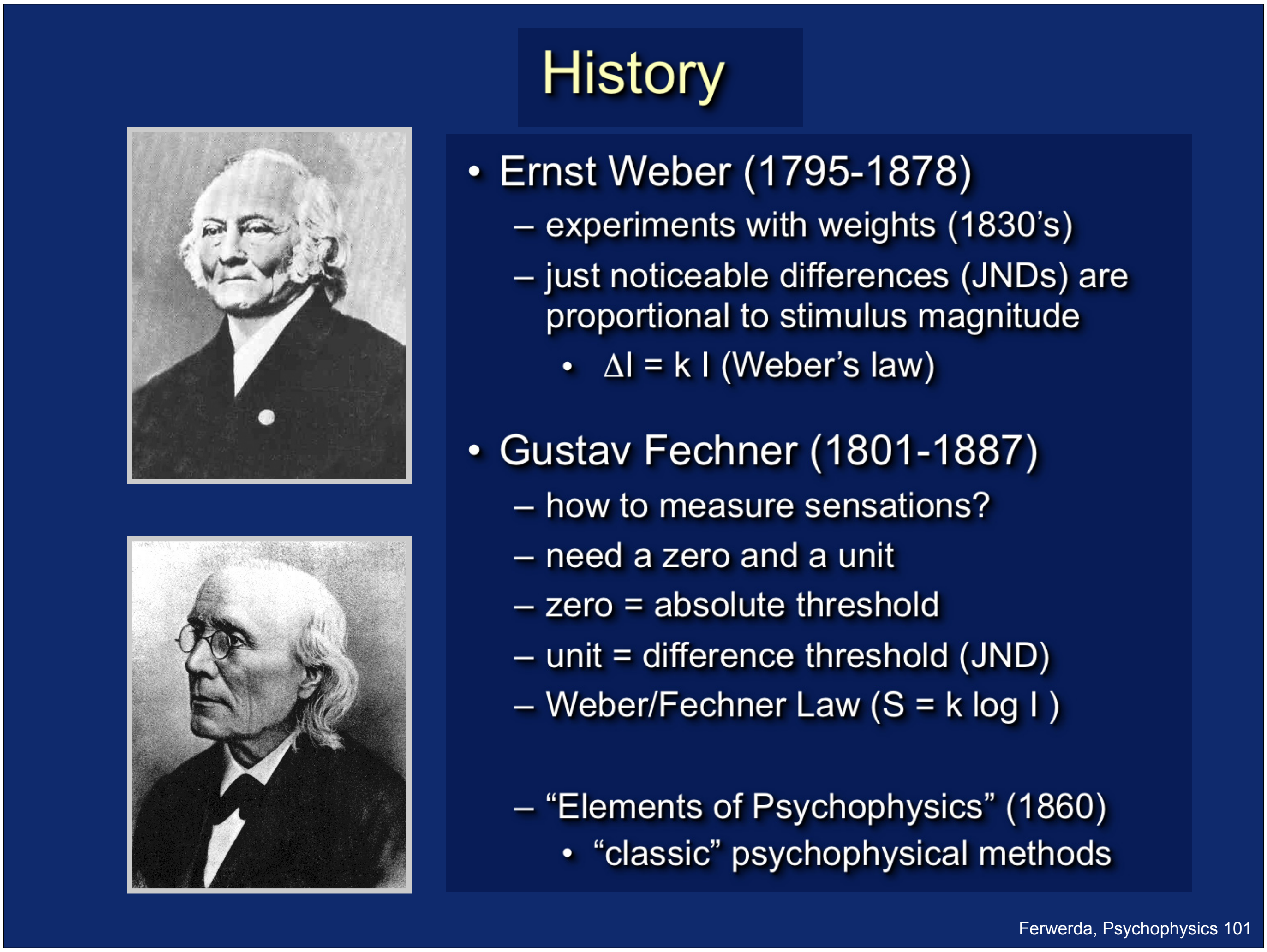
    examples: compression: JPEG, MPEG, MP3, graphics: visible difference predictors in image quality, perceptually-based rendering

- in addition to understanding absolute sensitivity and discrimination limits, we might want to know how what we perceive changes across some wide range of change in a physical parameter, this is the domain of scaling techniques
- in scaling interest is in quantifying the suprathreshold appearances of things, and in measuring how appearance changes with changes in physical properties

    for example: which square on the bottom is twice as bright as the one on the top

- scaling studies are also widely used in engineering application

    examples: Munsell color system, image quality metrics



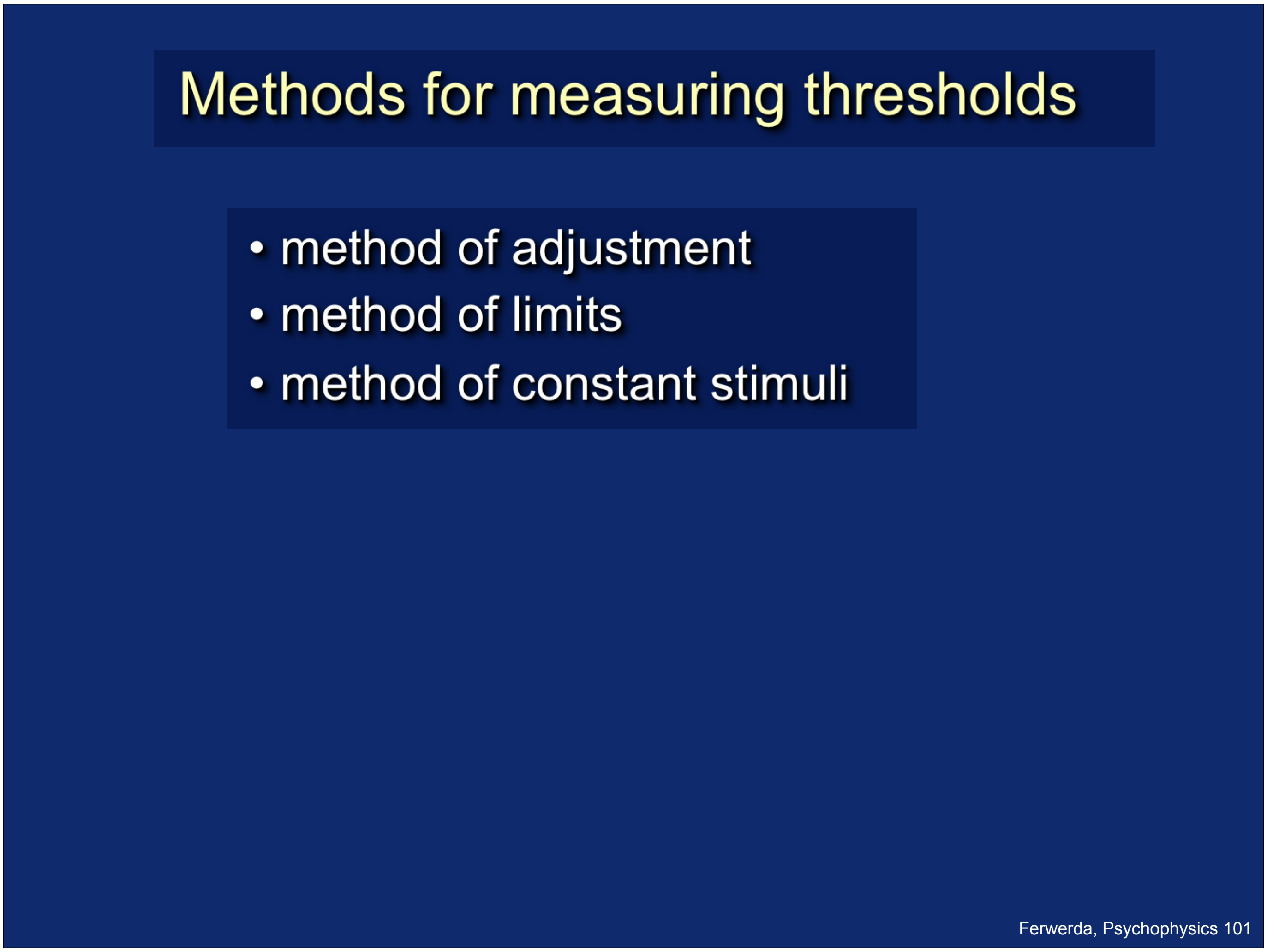
History

- Ernst Weber (1795-1878)
  - – experiments with weights (1830’s)
  - – just noticeable differences (JNDs) are proportional to stimulus magnitude
    - $\Delta I = k I$ (Weber’s law)
- Gustav Fechner (1801-1887)
  - – how to measure sensations?
  - – need a zero and a unit
  - – zero = absolute threshold
  - – unit = difference threshold (JND)
  - – Weber/Fechner Law ($S = k \log I$)
  - – “Elements of Psychophysics” (1860)
    - “classic” psychophysical methods

- development of psychophysical methods goes back over 150 years
- earliest work attributed to Weber
- experimented with weights, had subjects lift weights and judge differences
- found that just-noticeable-differences (JNDs) were proportional to size of the weight (i.e. needed greater differences to tell heavier weights apart)
- observations became formalized as Weber’s law which says
- delta I (JND) = k * I (magnitude of the standard)
- different “k’s” for different sensory modalities/judgments

- Fechner regarded as the father of psychophysics: relations between physical and psychological worlds
- basic question: how can we objectively measure subjective sensations?
- to measure anything need a zero and a unit value, then can build a ruler and do objective measurements
- Fechner took zero to be the absolute threshold (starting point of perception)
- drawing on Weber’s work, he took JND’s as the unit of perception, and by measuring JND’s at a range of physical values he could build up psychophysical scales to relate stimulation and perception
- because of the proportionality of JNDs, scales were all logarithmic, formalized as Fechner’s law
  - says sensation is proportional to the log of the physical stimulation, different k’s for different modalities, judgments

- Fechner’s lasting contribution was the development of a set of experimental methods for measuring psychophysical relations
- Published in “Elements of Psychophysics” in 1860

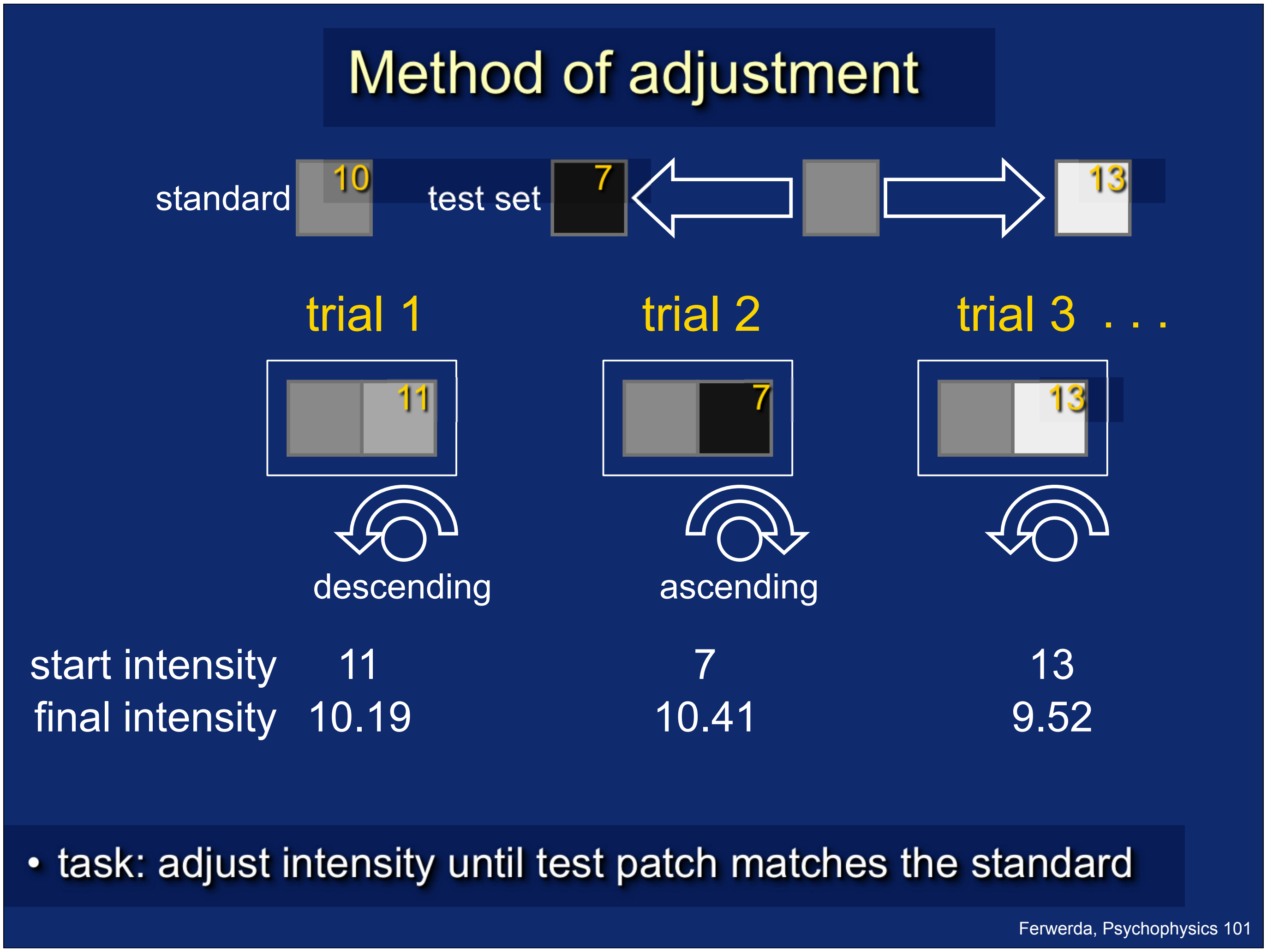


Methods for measuring thresholds

- method of adjustment
- method of limits
- method of constant stimuli

- rest of talk divided into two sections
- first section on techniques for measuring thresholds
- later will focus on techniques for developing psychophysical scales

- Fechner introduced three experimental methods for measuring psychophysical threholds
  - method of adjustment
  - method of limits
  - method of constant stimuli



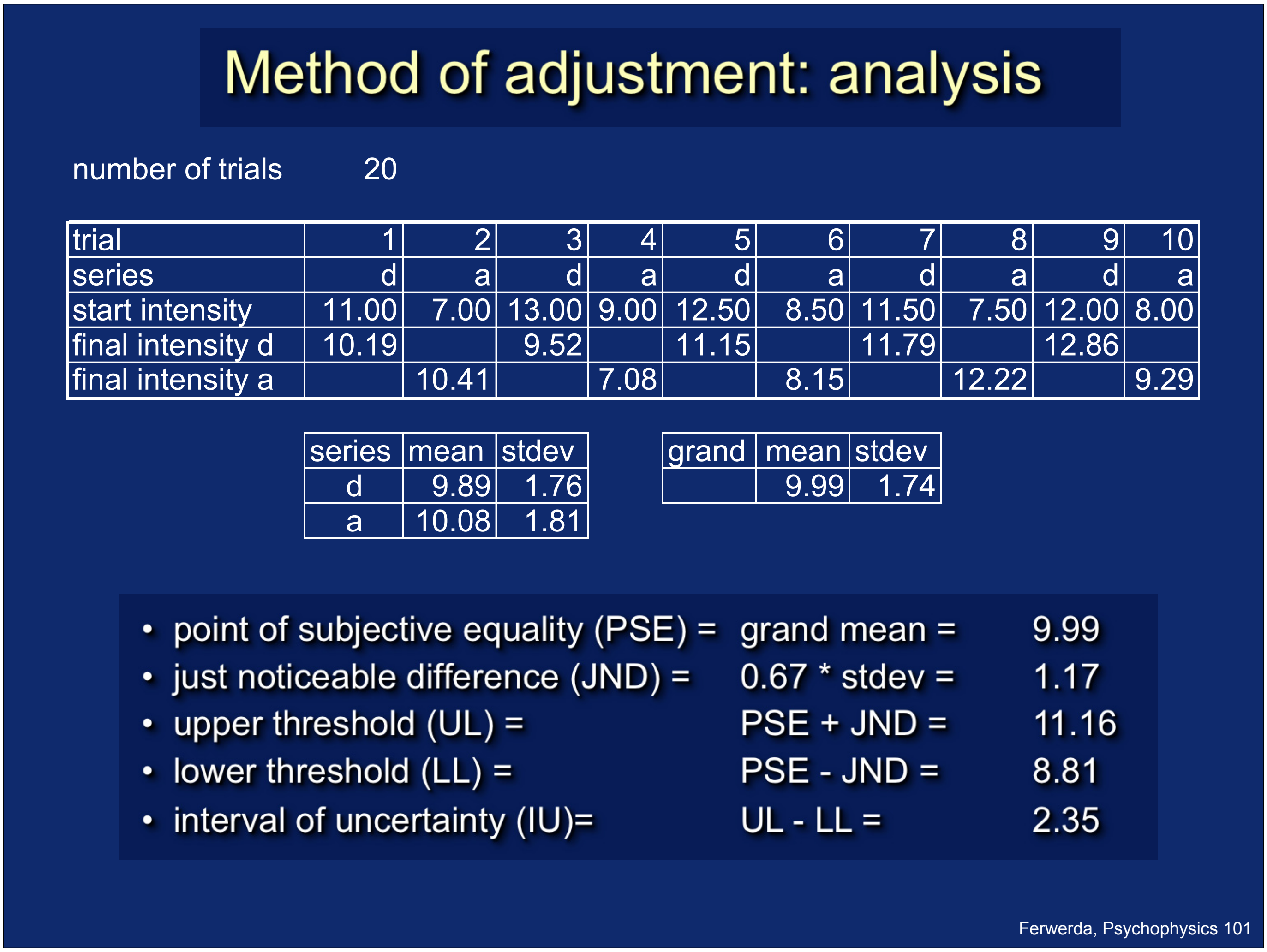
Method of adjustment

start intensity	11	7	13
final intensity	10.19	10.41	9.52

- task: adjust intensity until test patch matches the standard

Ferwerda, Psychophysics 101

- method of adjustment (MOA)
- example of measuring difference threshold for brightness
- start with fixed standard, continuously variable test set
- experiment is run as a series of trials
- on each trial standard and test are presented together (side-by-side, sequentially, or binocularly)
- subject's task: adjust intensity of the test square until just visibly different from standard
  - to control potential observer bias
    - starting intensity varied randomly
    - both descending trials (start intensity above standard) and ascending trials (start intensity below standard) are presented
- final intensities set are recorded for each trial



Method of adjustment: analysis

number of trials 20

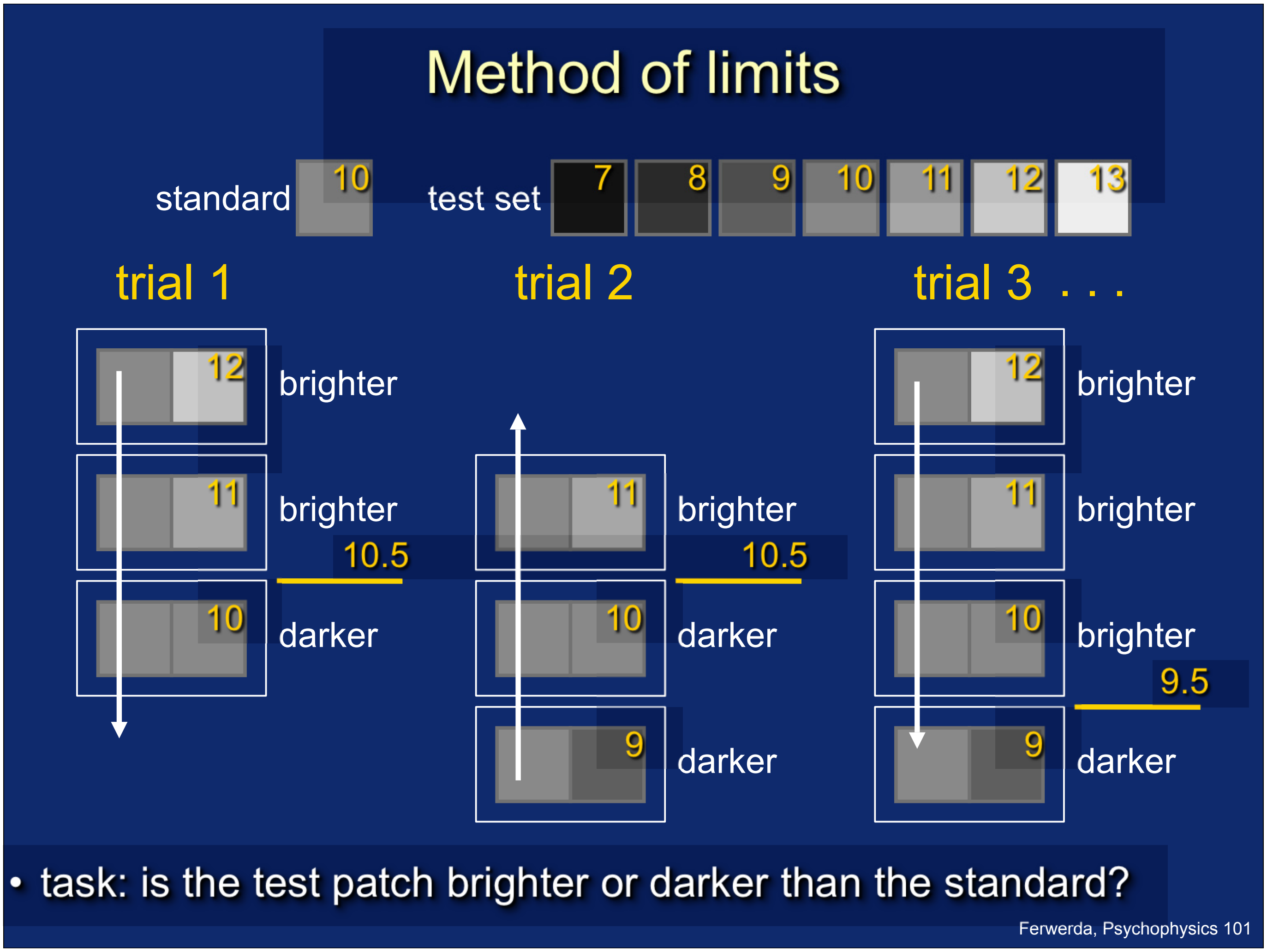
trial	1	2	3	4	5	6	7	8	9	10
series	d	a	d	a	d	a	d	a	d	a
start intensity	11.00	7.00	13.00	9.00	12.50	8.50	11.50	7.50	12.00	8.00
final intensity d	10.19		9.52		11.15		11.79		12.86	
final intensity a		10.41		7.08		8.15		12.22		9.29

series	mean	stdev
d	9.89	1.76
a	10.08	1.81

grand	mean	stdev
	9.99	1.74

- point of subjective equality (PSE) = grand mean = 9.99
- just noticeable difference (JND) = 0.67 * stdev = 1.17
- upper threshold (UL) = PSE + JND = 11.16
- lower threshold (LL) = PSE - JND = 8.81
- interval of uncertainty (IU)= UL - LL = 2.35

- to determine the threshold, we first tabulate the data (trials, series, start intensity, final intensity for ascending,descending)
- then calculate separate means and standard deviatons of the set values for the ascending and descending series
- check for bias by making sure that means are not too different from each other using a t-test
- If ok, can combine data and calculate overall mean and stdev for all trials
- we can use these statistics to derive several threshold values
  - point of subjective equality (PSE) – value where standard and test appear to match (usually physically equal)
  - just-noticeable-difference (JND) – proportional to the standard deviation of the responses, talk later about why particular mulitplier
  - two difference thresholds
    - upper – value where test is just noticeably brighter
    - lower – value where test is just noticeably darker
  - Interval of uncertainty – range where subject is uncertain whether test is brighter or darker
    - can be used as measure of how much precision required in representation
      - example: quantization in JPEG, as long as differences are within interval of uncertainty images will be “indistinguishable”

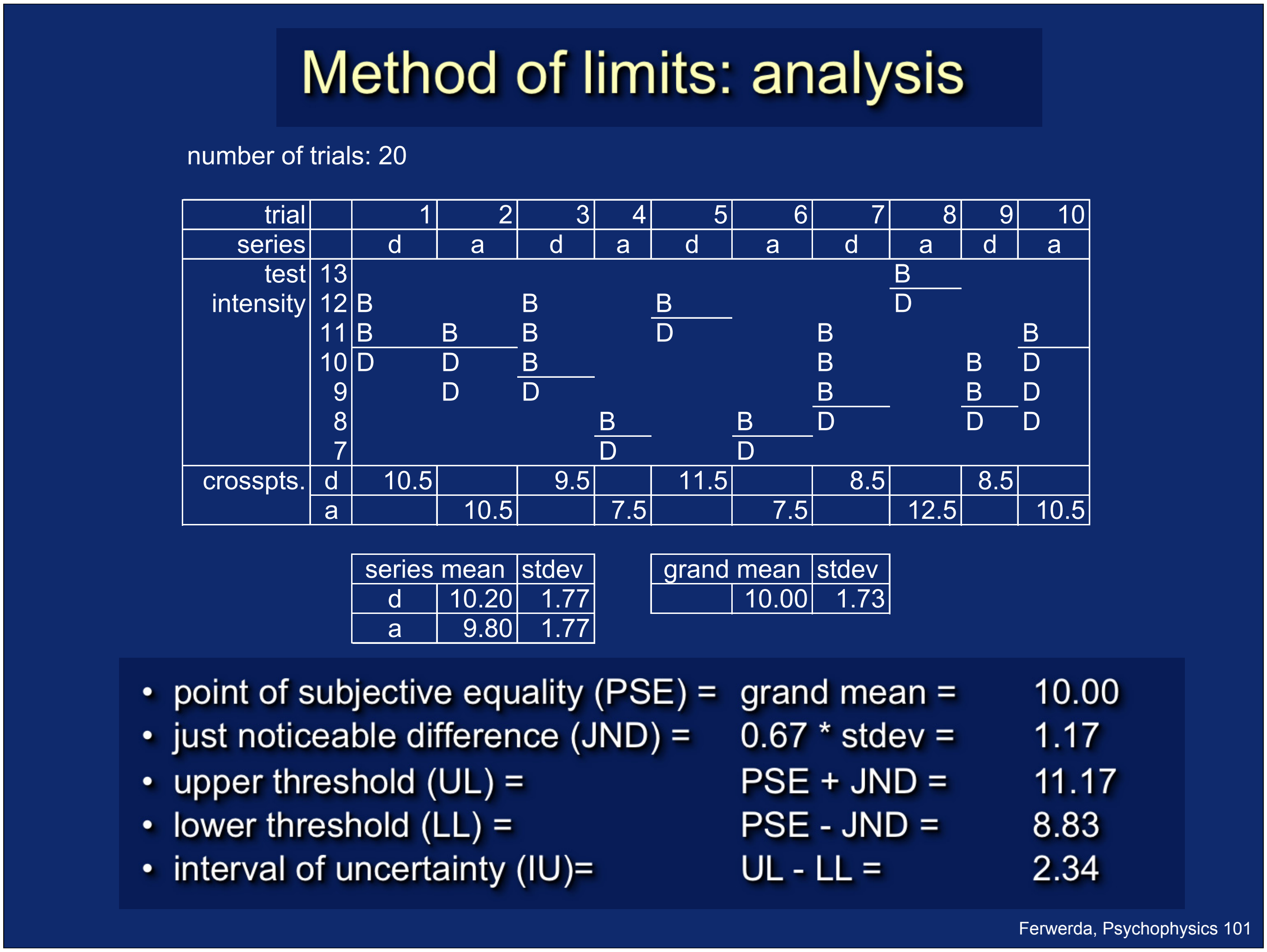


Method of limits

- task: is the test patch brighter or darker than the standard?

Ferwerda, Psychophysics 101

- MOA is fast, easy/interactive, but requires continuously variable stim, not always possible
- method of limits (MOL) addresses this, essentially a discrete stimulus version of MOA
- will use the same example: determining brightness difference thresholds
- fixed standard, test stimulus set with range that spans presumed threshold
- experiment is run as a series of trials
- on each trial standard and test presented together, subject asked if test is brighter or darker than standard
- if judged brighter, next darker stimulus is shown and the question is asked again
- trial proceeds until response changes, on change, value midway between brighter/darker appearing stimuli is recorded
- procedure similar to method of adjustment in terms of presenting a pattern of ascending, descending trials, random starting values



Method of limits: analysis

number of trials: 20

trial		1	2	3	4	5	6	7	8	9	10
series		d	a	d	a	d	a	d	a	d	a
test intensity	13								B		
	12	B		B		B			D		
	11	B	B	B		D		B			B
	10	D	D	B				B		B	D
	9		D	D				B		B	D
	8				B		B	D		D	D
	7				D		D				
crosspts.	d	10.5		9.5		11.5		8.5		8.5	
	a		10.5		7.5		7.5		12.5		10.5

series mean		stdev
d	10.20	1.77
a	9.80	1.77

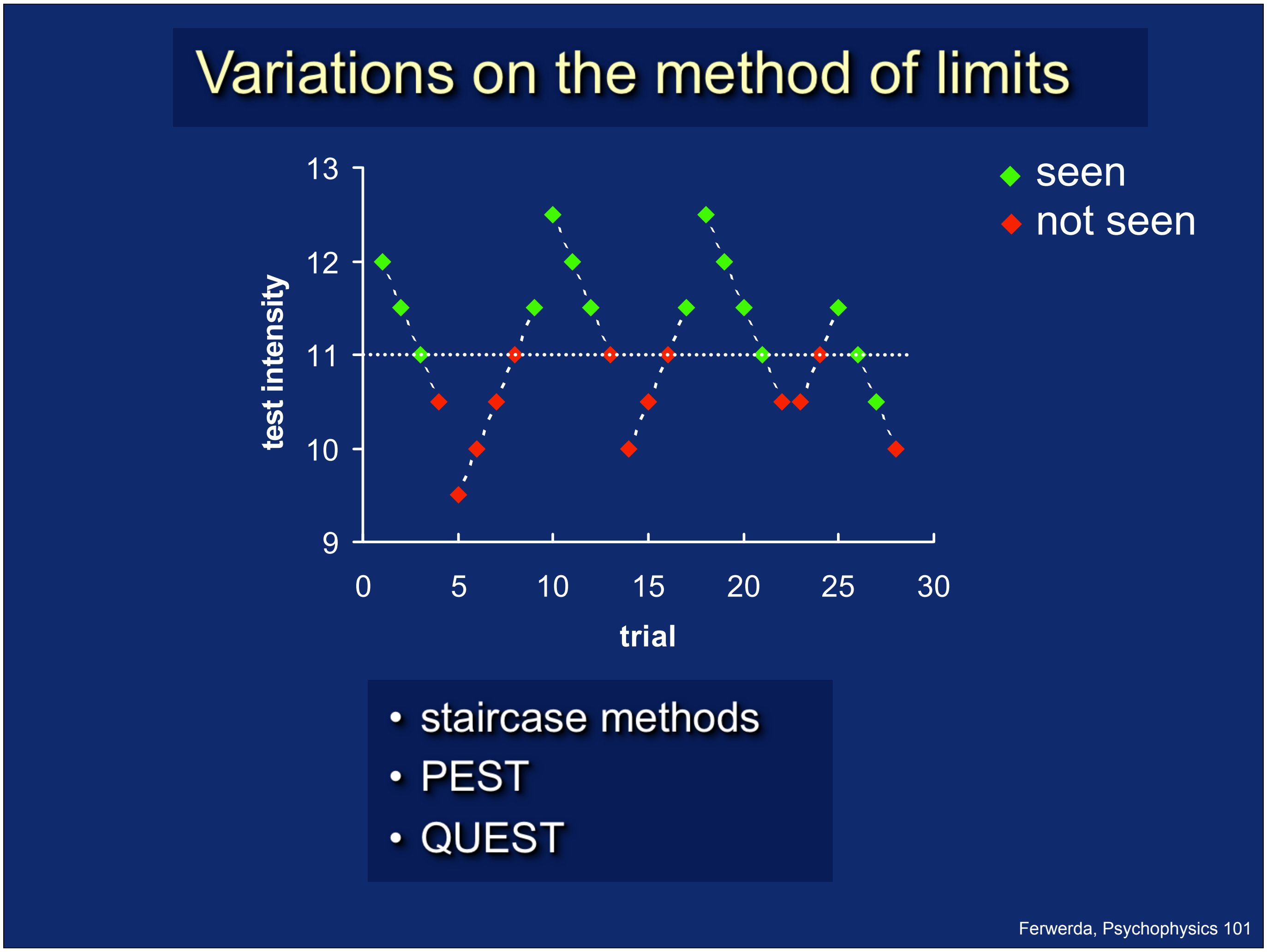
grand mean		stdev
	10.00	1.73

- point of subjective equality (PSE) = grand mean = 10.00
- just noticeable difference (JND) = 0.67 * stdev = 1.17
- upper threshold (UL) = PSE + JND = 11.17
- lower threshold (LL) = PSE - JND = 8.83
- interval of uncertainty (IU)= UL - LL = 2.34

Ferwerda, Psychophysics 101

- data is analyzed similarly to MOL
- responses to each test intensity are recorded for each descending and ascending trial series
- response crossover points are determined
- series and overall means and stdev’s are calculated and the various threshold values described earlier are determined

- advantage of method of limits: works for discrete stimulus sets
- disadvantage: inefficient method for finding threshold, requires many trials, lead to subject boredom, fatigue

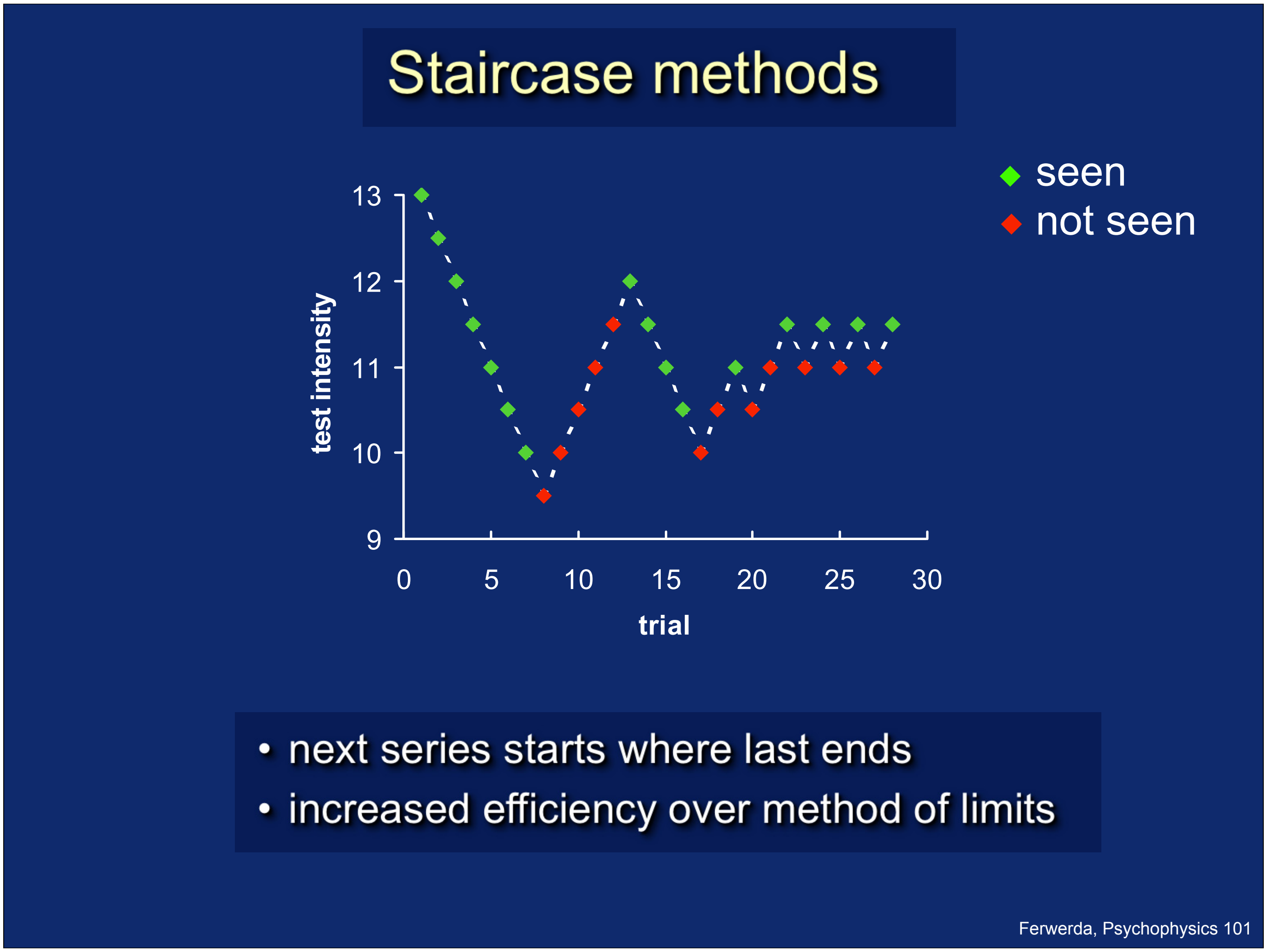


Variations on the method of limits

- staircase methods
- PEST
- QUEST

Ferwerda, Psychophysics 101

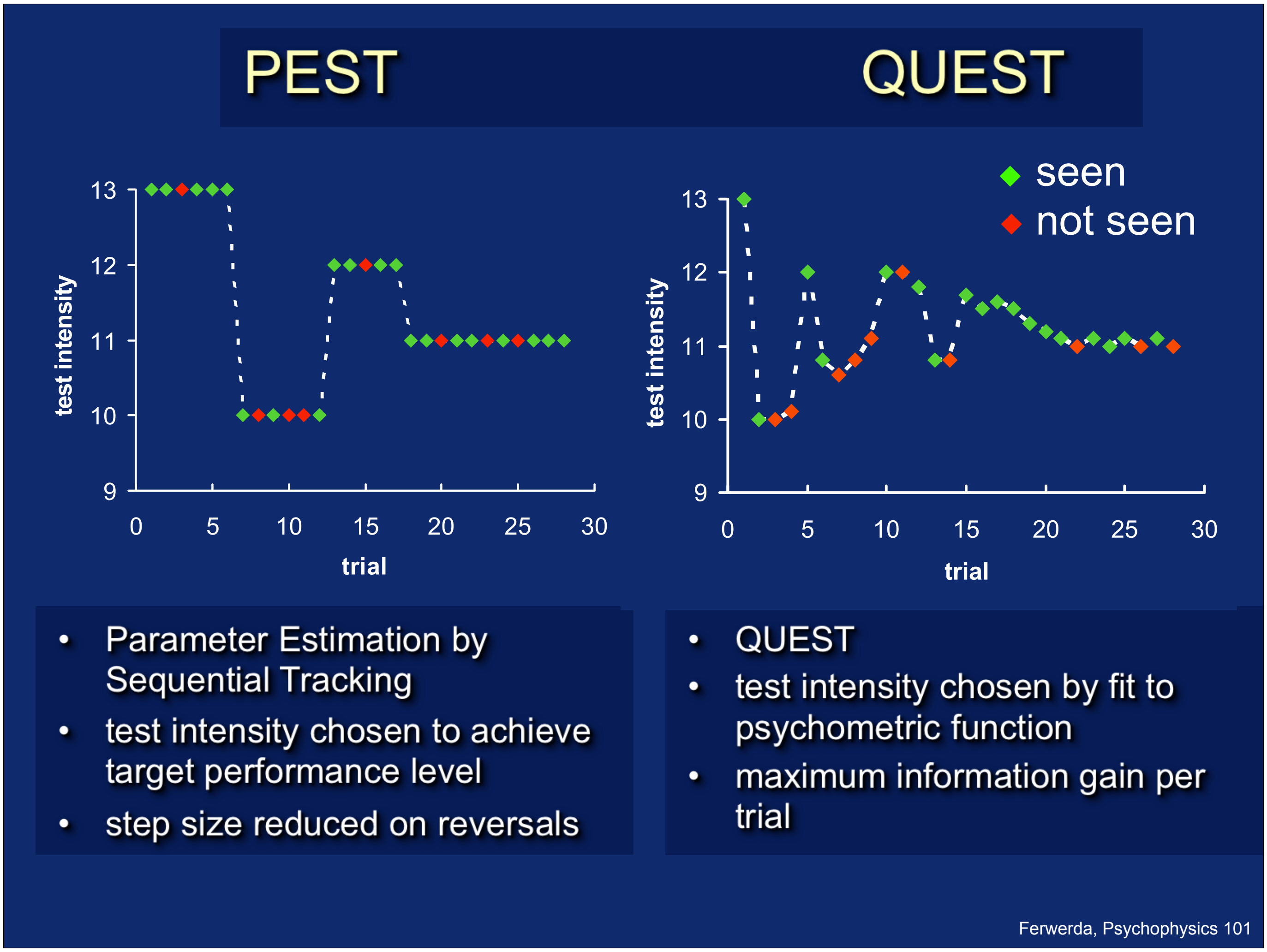
- to address efficiency issues variations on the MOL have been developed
  - staircase, PEST, QUEST
- graph shows a series of method of limits trials
- trial series repeatedly traverse threshold
- trials far from threshold don’t provide any information, waste time, effort



Staircase methods

- next series starts where last ends
- increased efficiency over method of limits

- one way to increase efficiency is to start next trial series where last one left off
- called staircase method because of step-up, step-down look of the trial record
- approach raises two issues 1) when to stop; 2) what values to include in threshold calculations
- different conventions used
  - stopping: after some number of reversals
  - data values: include all vs. discard first n “familiarization” trials



PEST QUEST

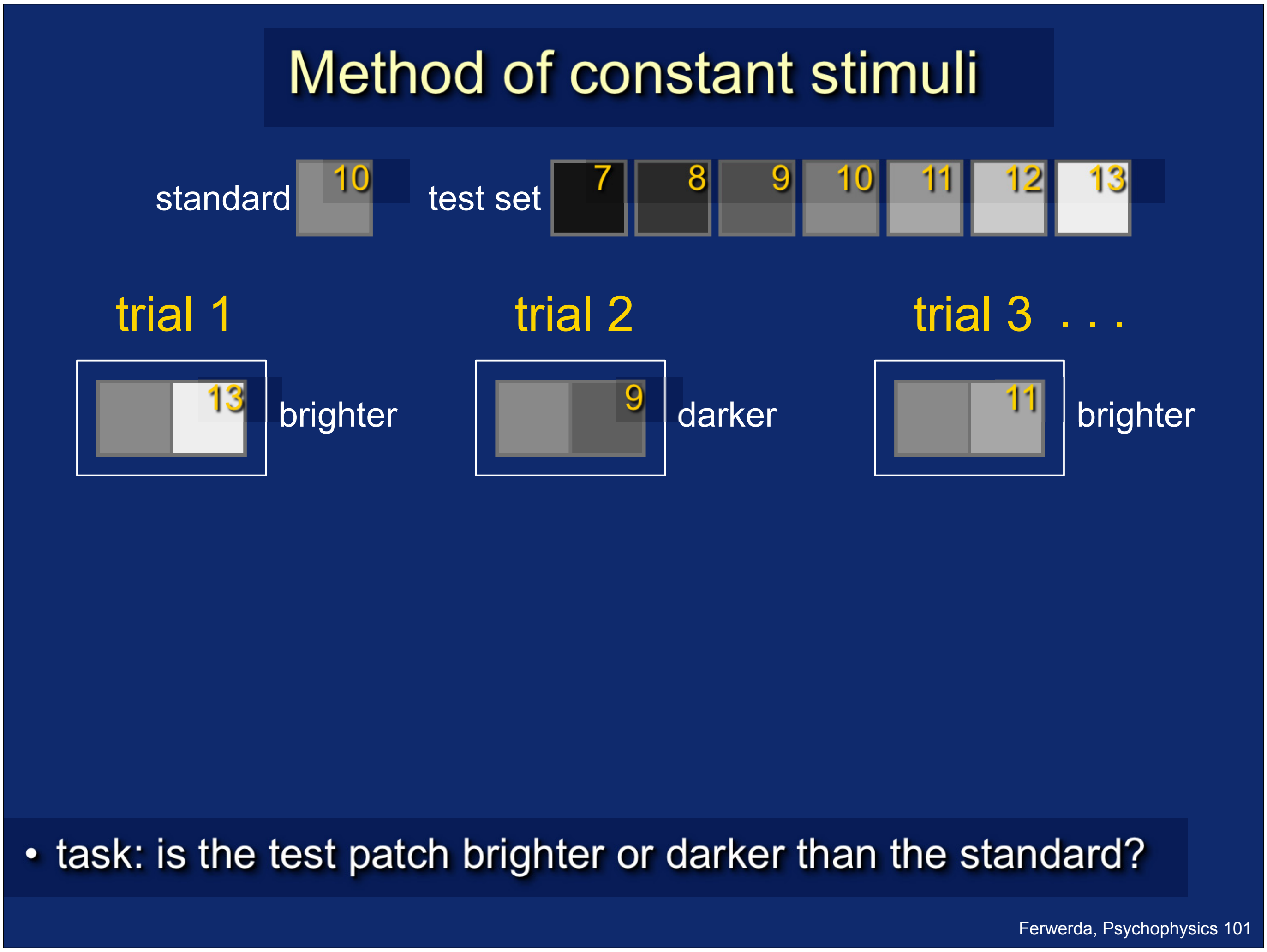
- Parameter Estimation by Sequential Tracking
- test intensity chosen to achieve target performance level
- step size reduced on reversals

- QUEST
- test intensity chosen by fit to psychometric function
- maximum information gain per trial

Ferwerda, Psychophysics 101

- staircase method improves the efficiency of the method of limits but stopping and data analysis criteria are somewhat ad-hoc
- In last 20 years two more principled improvements on the MOL have been introduced: PEST, QUEST
- PEST: Parameter Estimation by Sequential Tracking
- like staircase (step-up/down), but two differences
  - step size reduced on reversals
  - test intensity varied until a given performance level is achieved (e.g. 75% correct)

- other method is known as QUEST
- test intensity chosen so data gives best fit to what’s called a psychometric function
  - model of observer performance, will explain later
- particular advantage of QUEST is that it provides the maximum information gain per trial about the threshold value
  - increases efficiency: threshold can be determined in the shortest number of trials



Method of constant stimuli

standard 10 test set 7 8 9 10 11 12 13

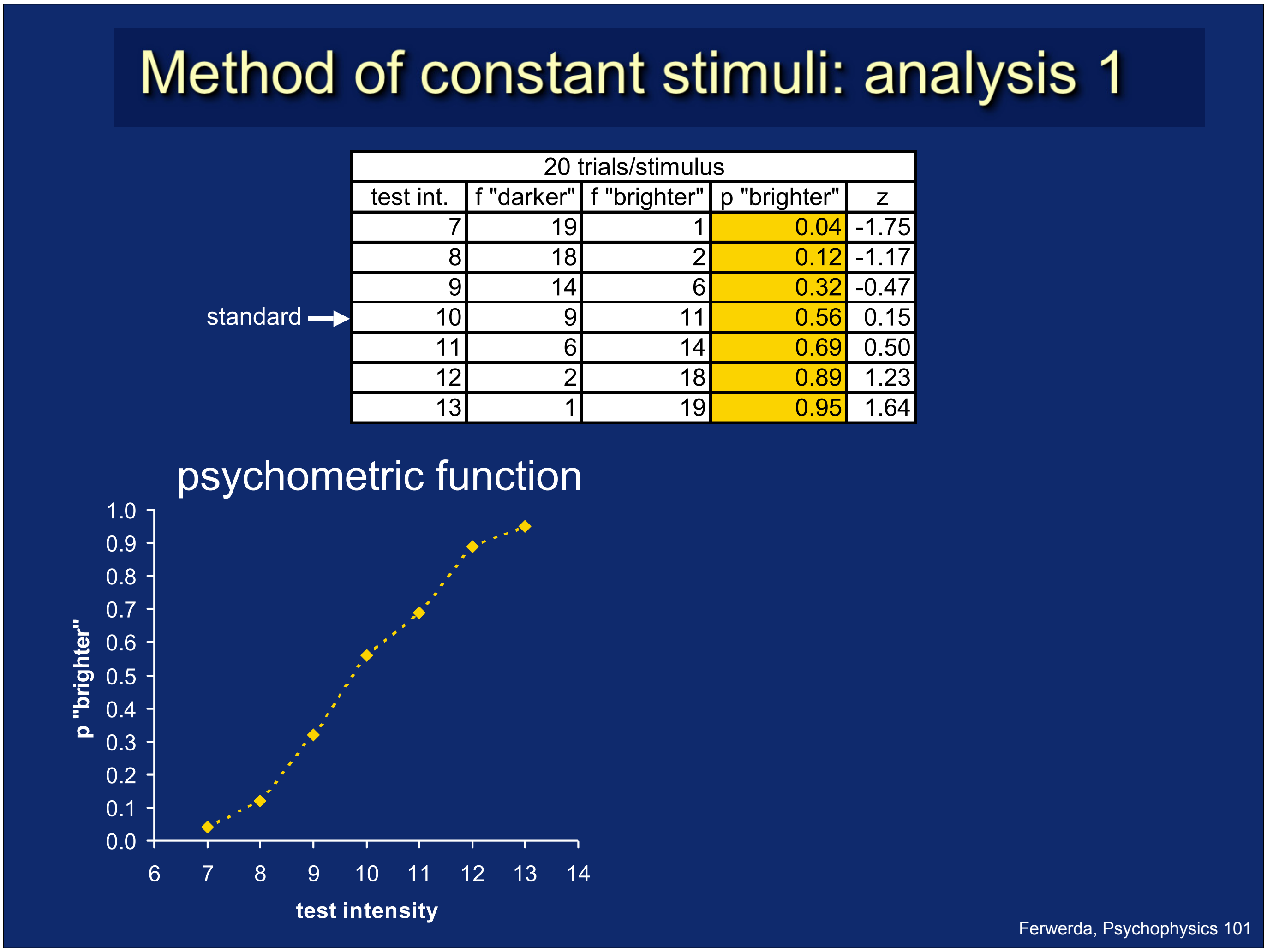
trial 1 13 brighter

trial 2 9 darker

trial 3 . . . 11 brighter

- task: is the test patch brighter or darker than the standard?

- third of the Fechner’s classical threshold methods is method of constant stimuli (MCS)
- essentially randomized version of the method of limits
- standard, discrete test set
- experimental method is simple: series of trials where members of the test set are presented with the standard in pairs
- subject is asked if test square is brighter or darker than the standard
- random presentation controls for a variety of bias/errors that occur with ordered presentations like MOL



Method of constant stimuli: analysis 1

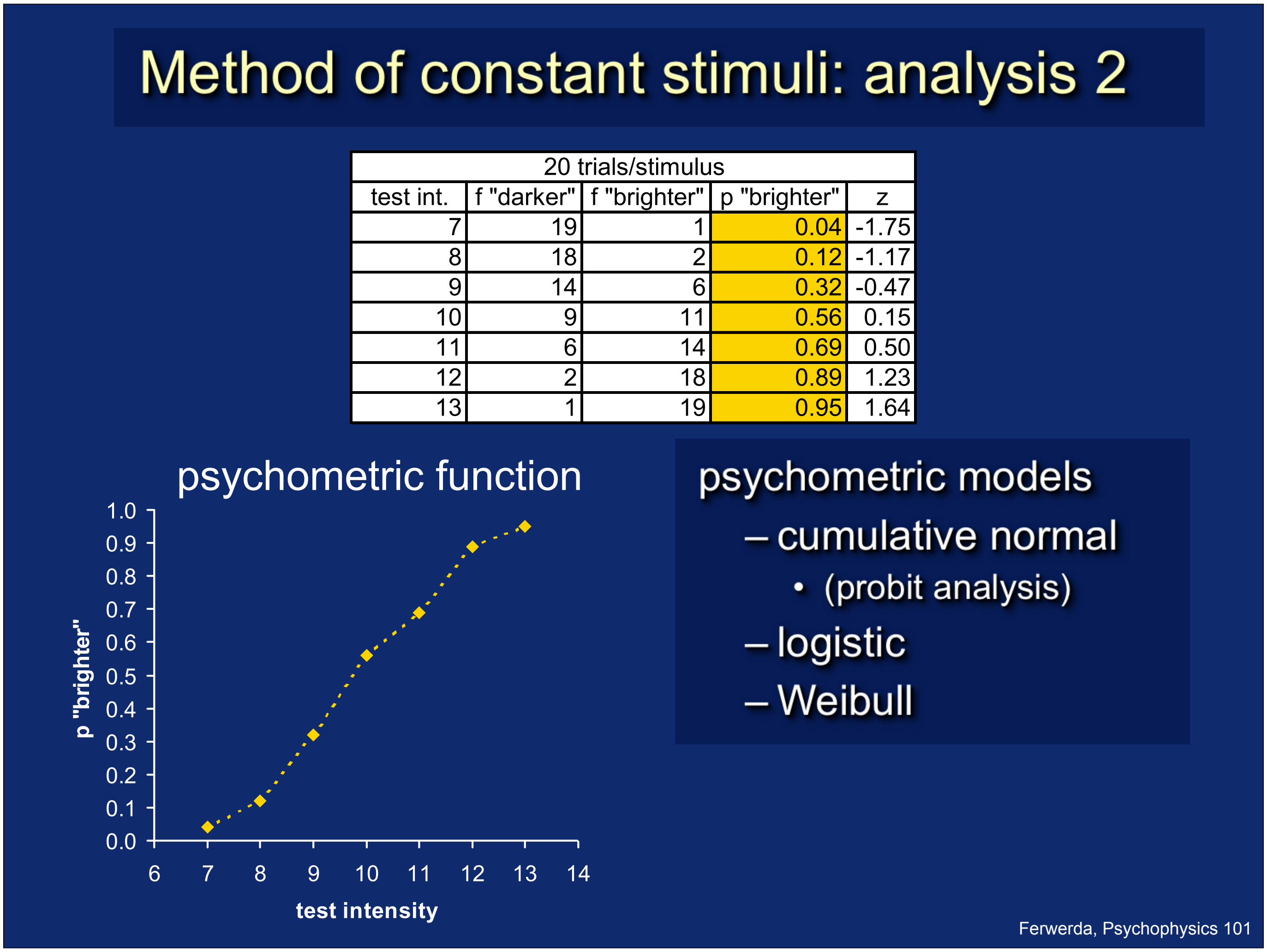
20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

standard → 10

psychometric function

Ferwerda, Psychophysics 101

- dat are analyzed by first tabulating how frequently each test patch was judged brighter or darker than the standard
- 20 trials/stim, standard intensity 10
- test intensity 7: darker 19, brighter 1
- test intensity 12: darker 2, brighter 18
- test intensity 10: darker 9, brighter 11 (uncertainty in response)
- can convert frequencies to proportions by dividing by number of trials (in this case 20)
- p's are the probability that a given test intensity will be seen as brighter than the standard
- if we plot p's vs. test intensity we get what is known as a psychometric function (or frequency of seeing curve)
- shows that the liklihood of seeing the test as brighter than the standard increases with the intensity of the standard
- but where is the threshold?



Method of constant stimuli: analysis 2

20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

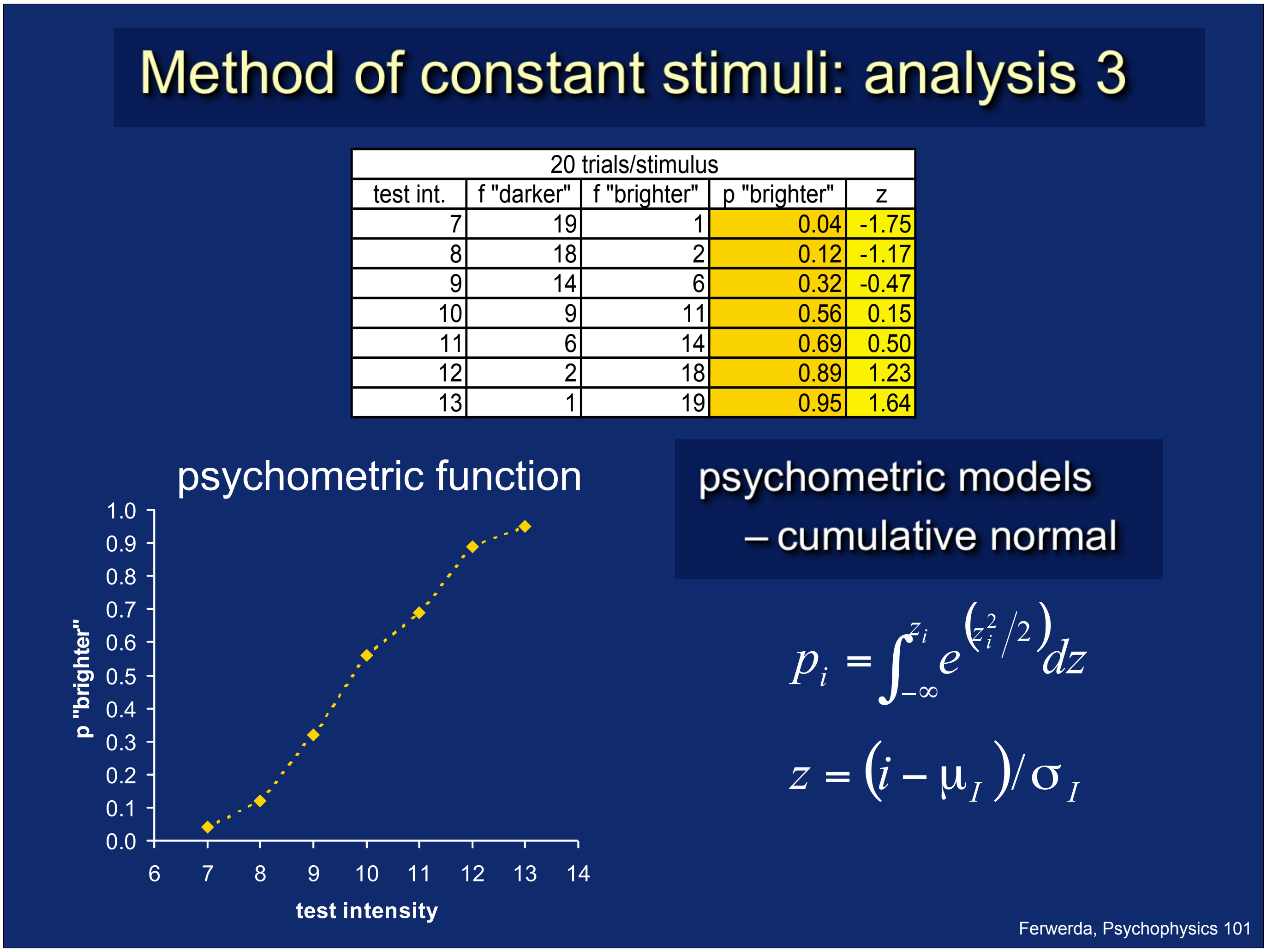
psychometric function

psychometric models

- cumulative normal
  - (probit analysis)
- logistic
- Weibull

Ferwerda, Psychophysics 101

- to determine the thresholds for discriminating between the patches, we need to fit the data with a model that describes the observer’s response behavior
- there are a number of different mathematical models:
  - cumulative normal (a.k.a. probit analysis)
  - logistic function
  - Weibull function
- subtle differences between these models, used by different communities, all basically S-shaped functions
- for illustration we’re going to use the cumulative normal model



Method of constant stimuli: analysis 3

20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

psychometric function

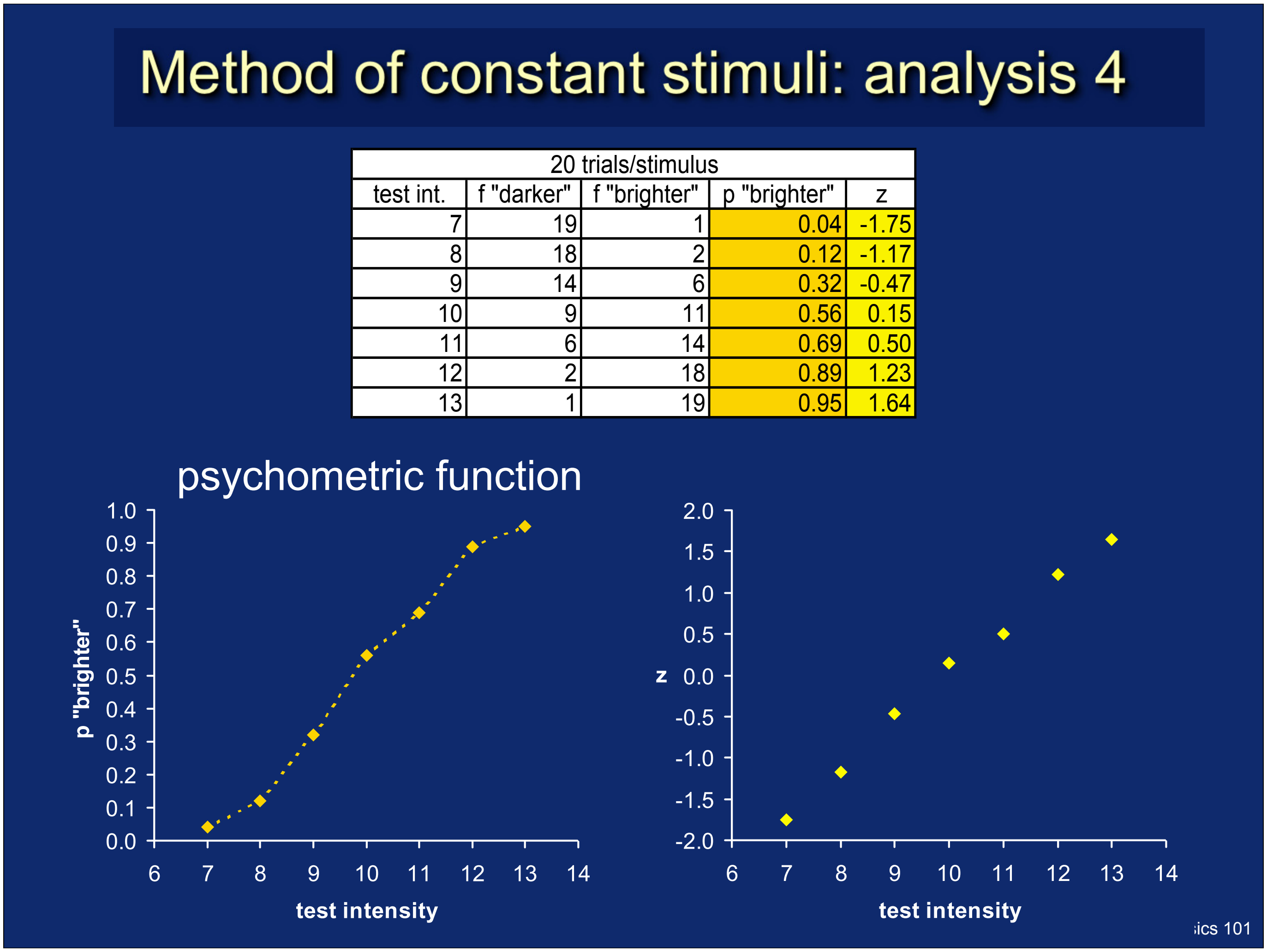
psychometric models

– cumulative normal

$$p_i = \int_{-\infty}^{z_i} e^{\left(z_i^2/2\right)} dz$$

$$z = \left(i - \mu_I\right)/\sigma_I$$

- function on the right shows the model
- p is the integral of an exponential function
- z is a parameter that is related to both the mean and the variance of the responses collected in the experiment

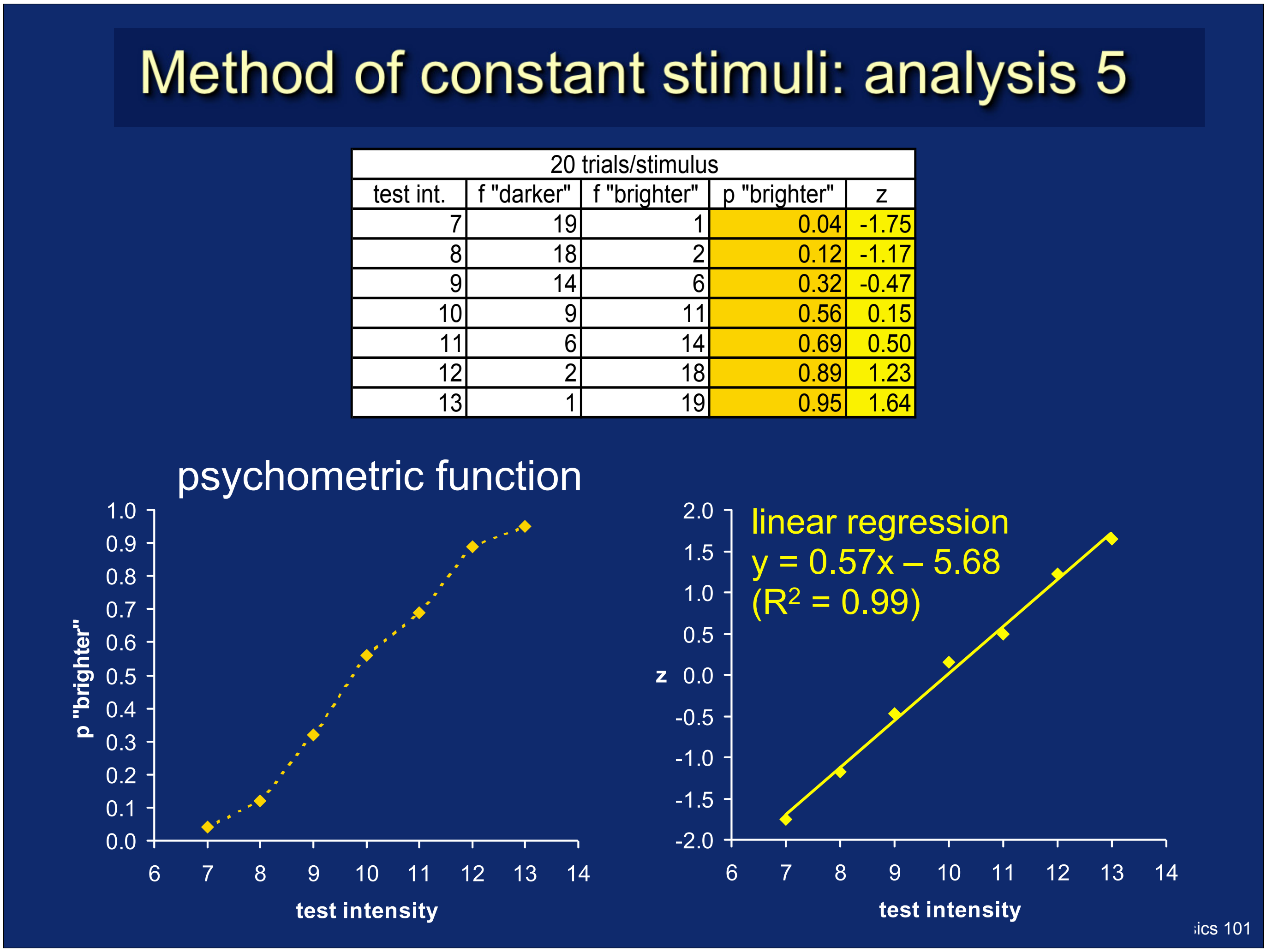


Method of constant stimuli: analysis 4

20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

psychometric function

- we can calculate the z values that correspond to the probabilities in the psychometric (cumulative normal) function (NORMSINV function in Excel)
- and we can plot these z values as a function of the test intensities



Method of constant stimuli: analysis 5

20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

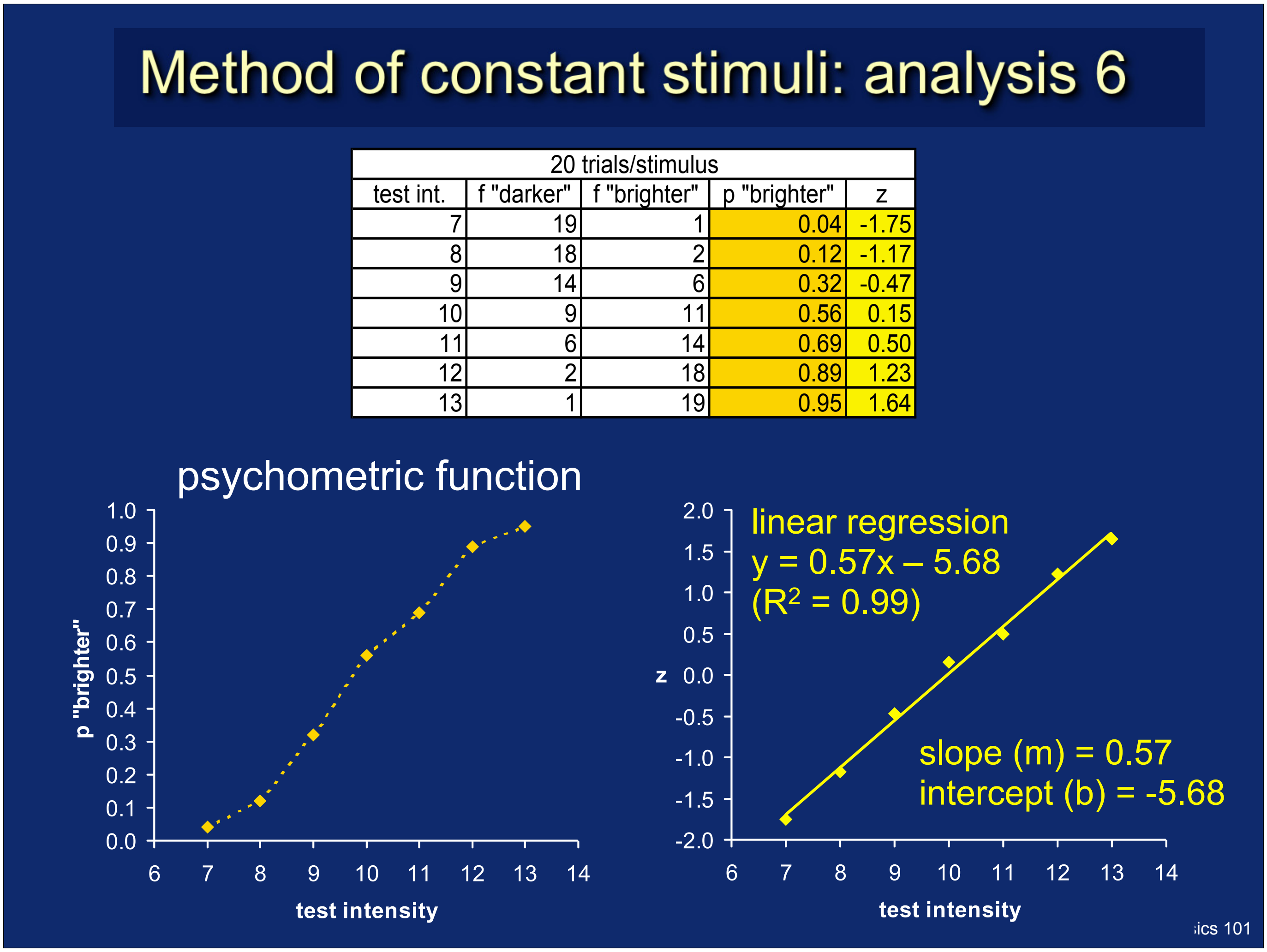
psychometric function

linear regression

$y = 0.57x – 5.68$

$(R^2 = 0.99)$

- if the cumulative normal model is a good fit to the data, then the z values should fall along a straight line
- by using linear regression (“fit trendline” in Excel) we can find both the slope and intercept of this line and the goodness of fit
- R^2 of 0.99 shows that the normal model is a good fit to the data



Method of constant stimuli: analysis 6

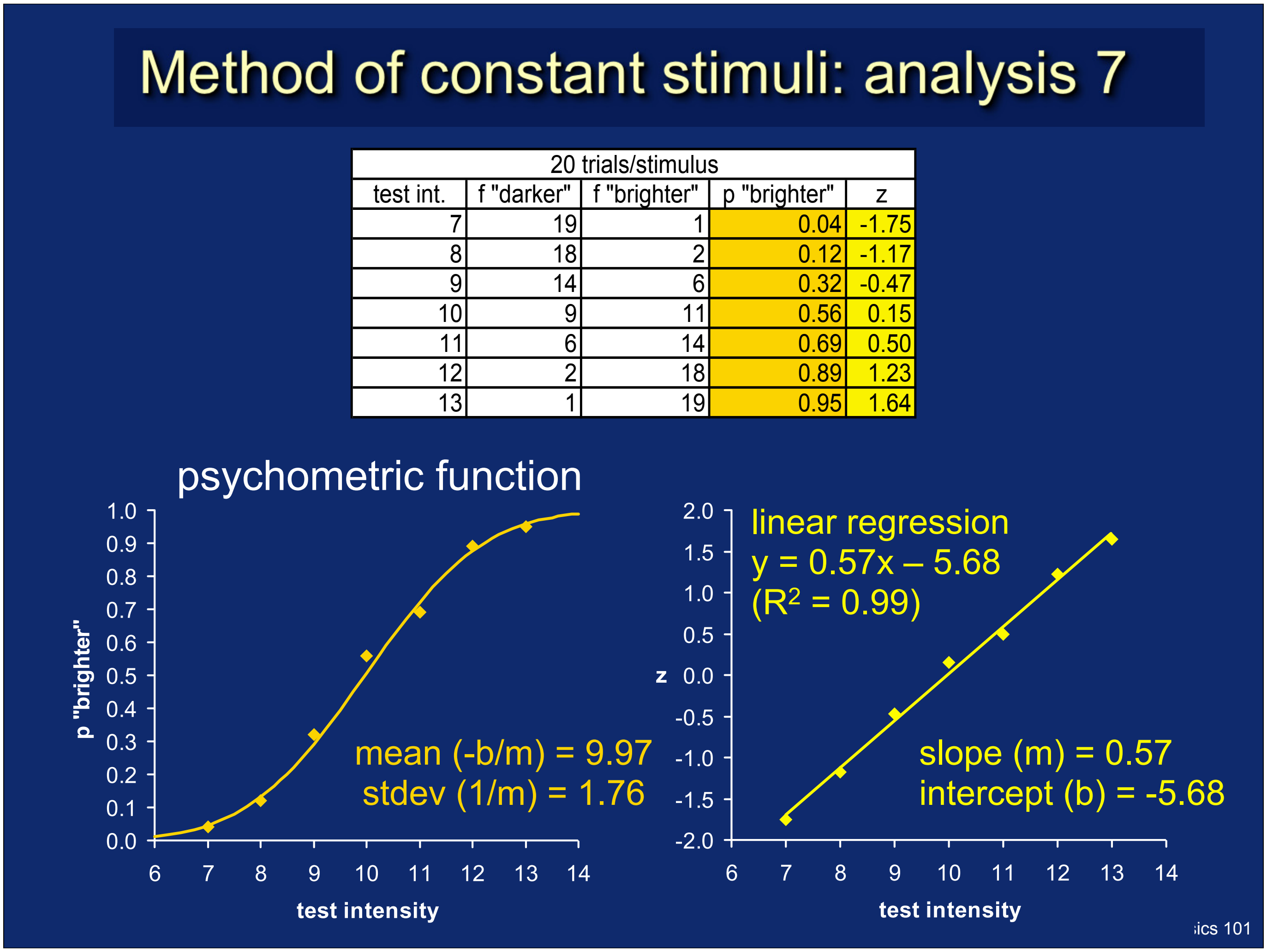
20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

psychometric function

linear regression
$y = 0.57x – 5.68$
$(R^2 = 0.99)$

slope (m) = 0.57
intercept (b) = -5.68

- slope (m) = 0.57
- intercept (b) = -5.68



Method of constant stimuli: analysis 7

20 trials/stimulus				
test int.	f "darker"	f "brighter"	p "brighter"	z
7	19	1	0.04	-1.75
8	18	2	0.12	-1.17
9	14	6	0.32	-0.47
10	9	11	0.56	0.15
11	6	14	0.69	0.50
12	2	18	0.89	1.23
13	1	19	0.95	1.64

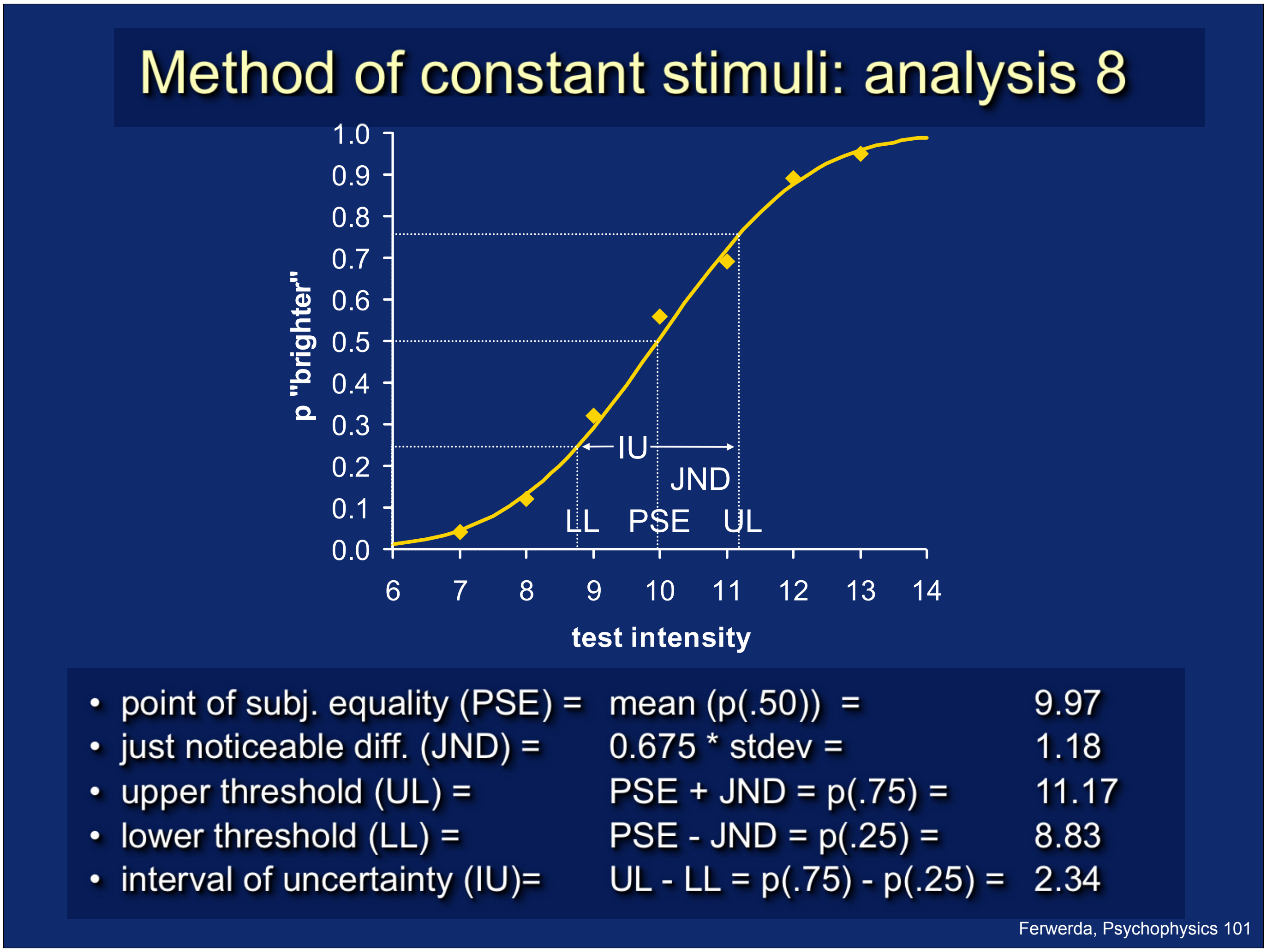
psychometric function

mean (-b/m) = 9.97
stdev (1/m) = 1.76

linear regression
y = 0.57x – 5.68
(R^2 = 0.99)

slope (m) = 0.57
intercept (b) = -5.68

- can use slope and intercept to calculate the mean and stdev of the cumulative normal function that fits the data
- graph on left shows this S-shaped curve and its fit to the data



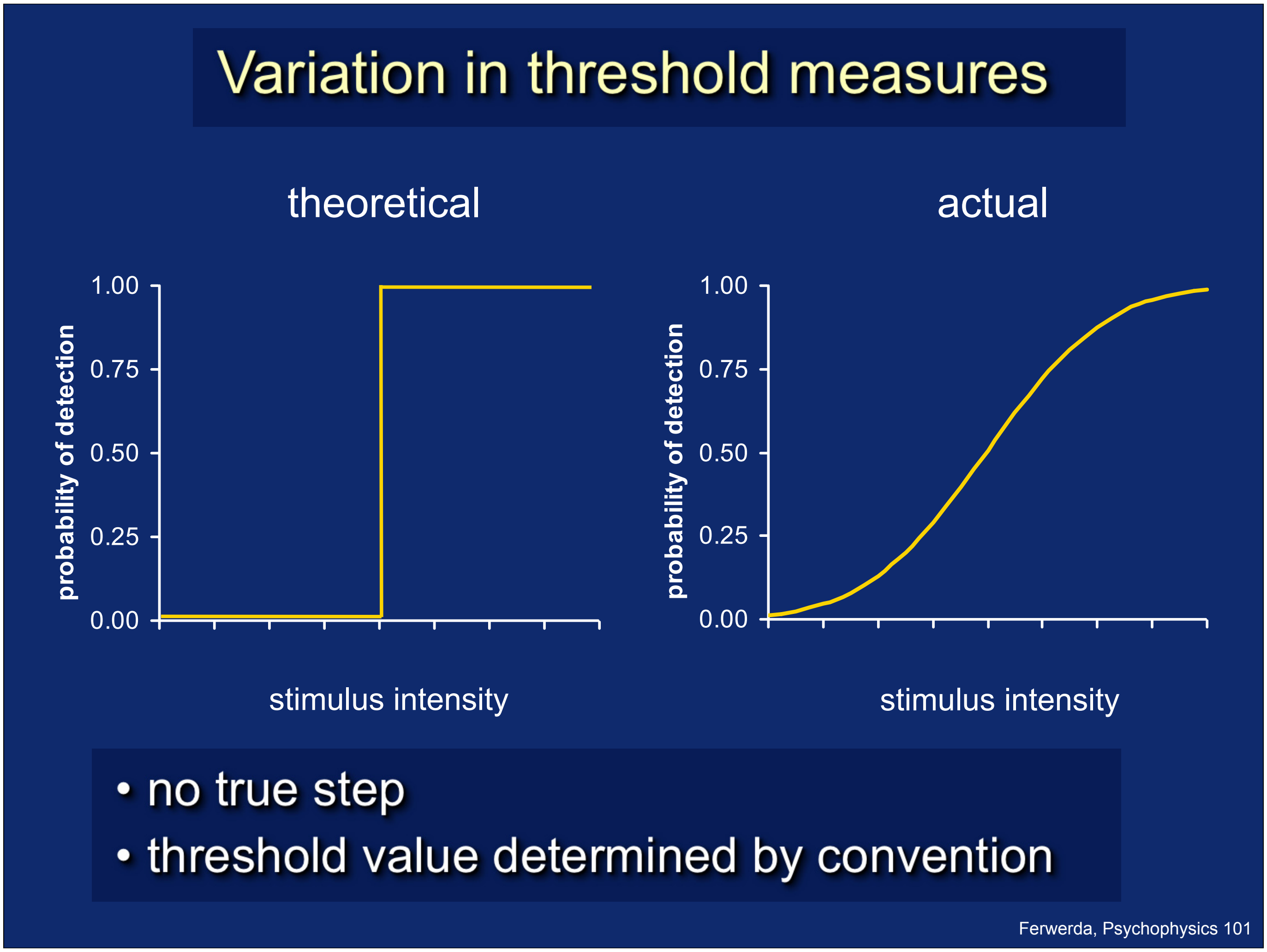
Method of constant stimuli: analysis 8

- point of subj. equality (PSE) = mean (p(.50)) = 9.97
- just noticeable diff. (JND) = 0.675 * stdev = 1.18
- upper threshold (UL) = PSE + JND = p(.75) = 11.17
- lower threshold (LL) = PSE - JND = p(.25) = 8.83
- interval of uncertainty (IU)= UL - LL = p(.75) - p(.25) = 2.34

Ferwerda, Psychophysics 101

- can use this model to determine the standard set of threshold values for the experiment
- point of subjective equality corresponds to the mean of the data (here given by (p= 0.50))
  - equal probability of seeing test patch as brighter or darker
  - value is close to the standard so no observer/experimental bias
- JND is again proportional to the standard deviation, but now we can see why the 0.67 factor is used:
  - because under the cumulative normal model adding this value to the PSE brings us to the (p=0.75) point on the curve (which by convention is taken as the upper threshold value)
  - reason for (p=.75): this is halfway between perfect performance (p=1.0) and the (p=0.5) performance that would be expected from just guessing
  - in practical terms p= 0.75 means that a test intensity of 11.17 will be seen as brighter than the standard on 3 out of 4 trials
- last two lines show that we can calculate the lower threshold and interval of uncertainty using similar logic

- point of this exercise: in any analysis of experimental data some statistical model is always used
- in the analyses of the first two methods we just assumed that a normal model fit the data and used means and standard deviations to calculate the threshold values
- but in this example we were actually explicit about testing whether the model we used (in this case the cumulative normal) was actually a good description of the data



Variation in threshold measures

- no true step
- threshold value determined by convention

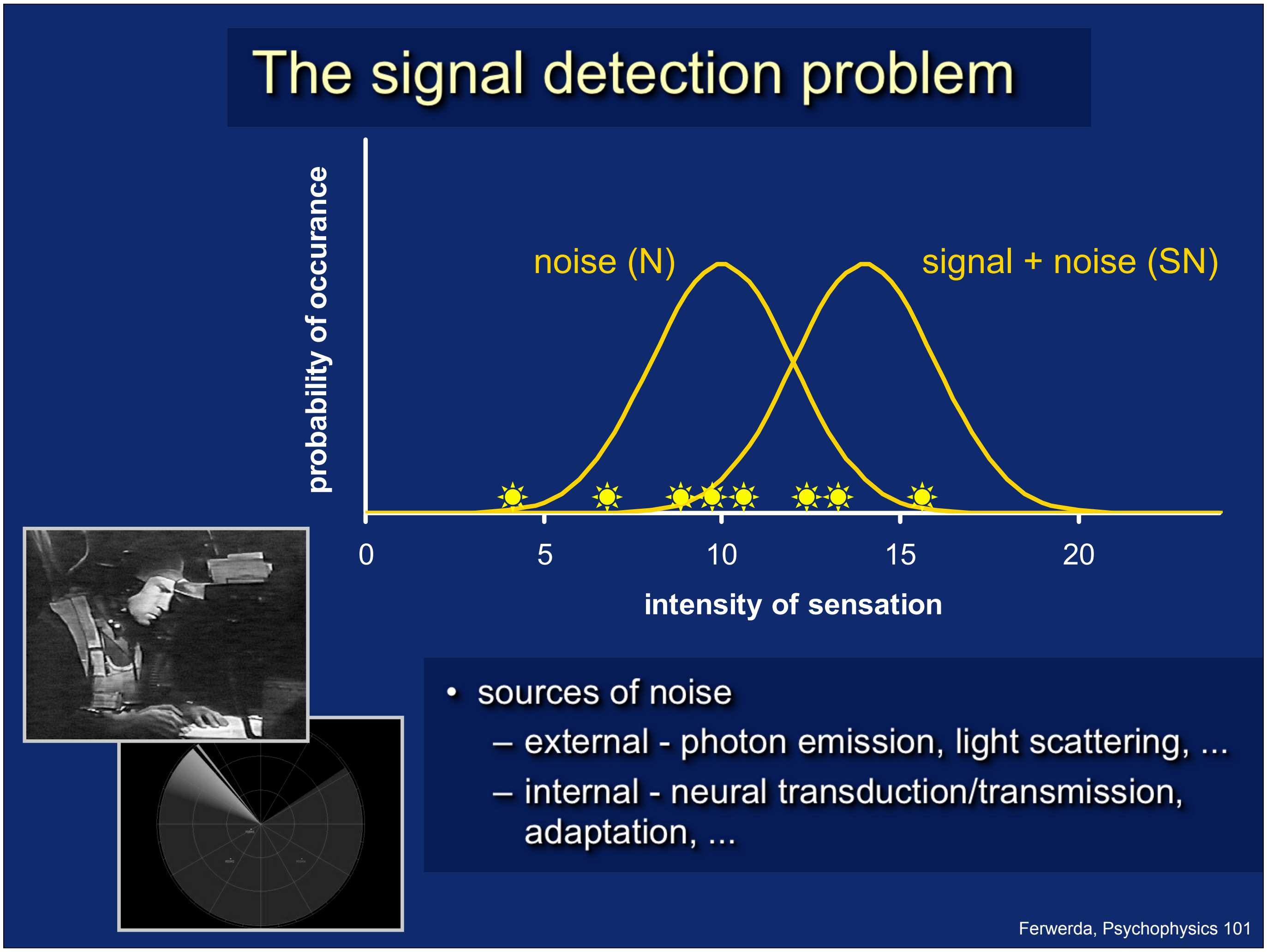
- usually think of threshold as a step/ breakpoint between seeing and not seeing
- shown in graph on left
- increase stimulus intensity, some low range where it’s undetectable (or indiscriminable) (probability 0.0)
- but at some point the threshold is crossed and the stimulus is always seen (probability (1.0)

- in reality we see that real observers perform differently
- actual threshold data shows that in many experiments the probability of detecting the stimulus increases with intensity
- but in reality there is no discrete step in the curve that corresponds to the theoretical notion of a threshold

- why do observers behave this way?
- what does this mean about our ideas of thresholds in perception?

The signal detection problem

- to get a handle on variation in threshold measures we need to understand what’s known as the signal detection problem
- motivation for Signal Detection Theory (SDT)
- SDT developed in the 40’s
- provides a framework for predicting observer’s behavior on important perceptual tasks
    - such as radar operator
    - radiologist, x-ray
    - scientific visualization, data mining applications where judgment is based on subtle visual cues



The signal detection problem

- sources of noise
  - – external - photon emission, light scattering, ...
  - – internal - neural transduction/transmission, adaptation, ...

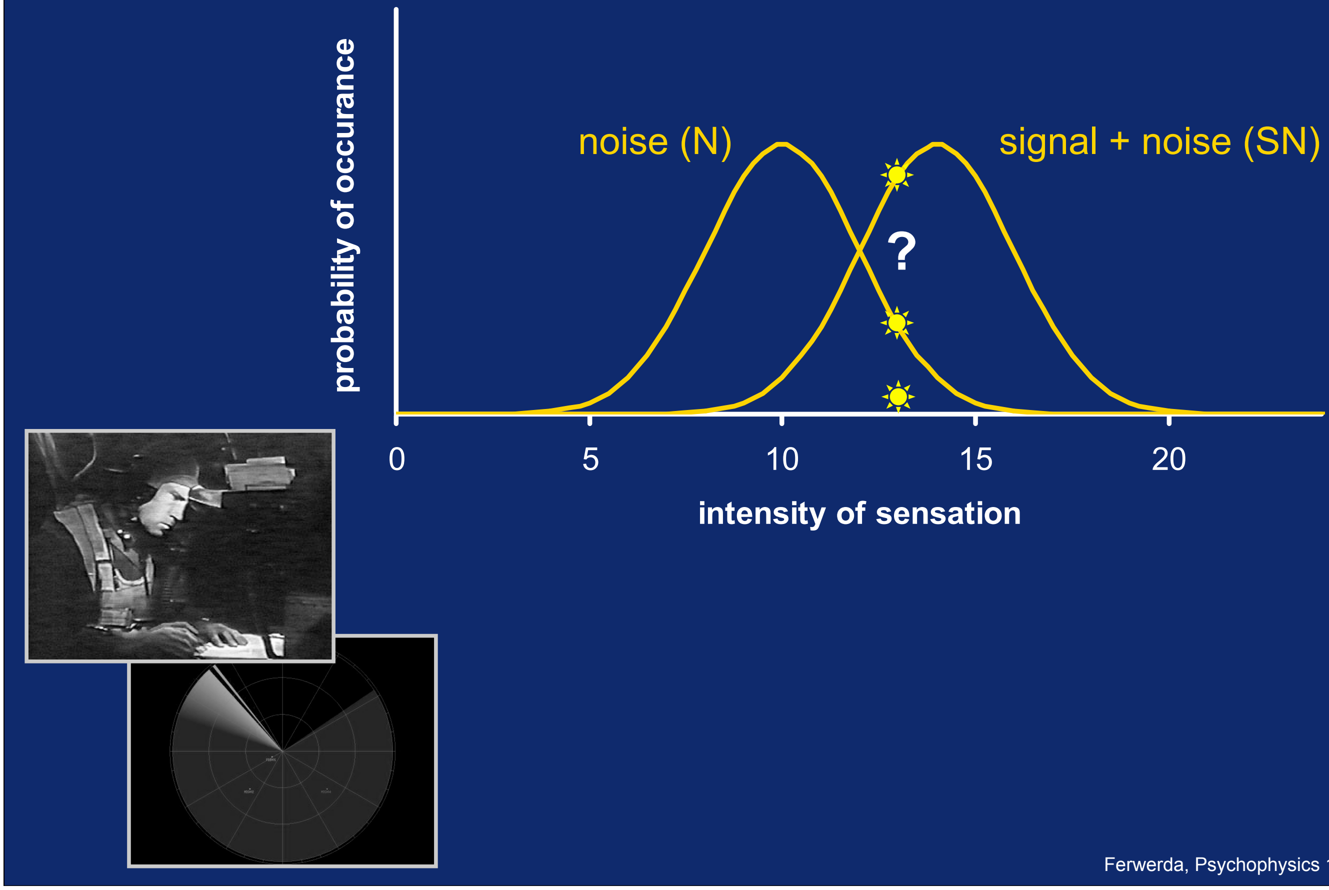
- original context
- imagine that you're a radar operator
- looking at screen, trying to see blips of light that represent enemy planes
- very important that you detect the planes, but very hard because the blips are sometimes faint

- so imagine you're focused on a particular location on the screen
- for each sweep of the beam (a trial) you're going to experience some intensity
- however because of noise, the value of the intensity that you'll experience will vary from sweep to sweep
- both external (physical) and internal (neural/psychological)sources of noise

- over time, the probability that you'll experience a given intensity of sensation on a trial can be described by the noise distribution (N)
- (N) represents the variation in perceived intensity just due to noise in the system

- applying a fixed intensity signal as input to the system will shift distribution up because of noise, range of sensations will remain the same (SN distribution)

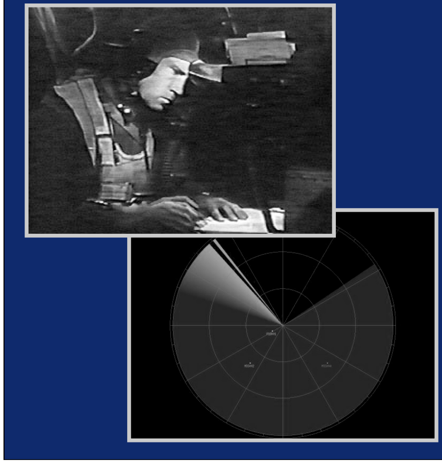
The signal detection problem



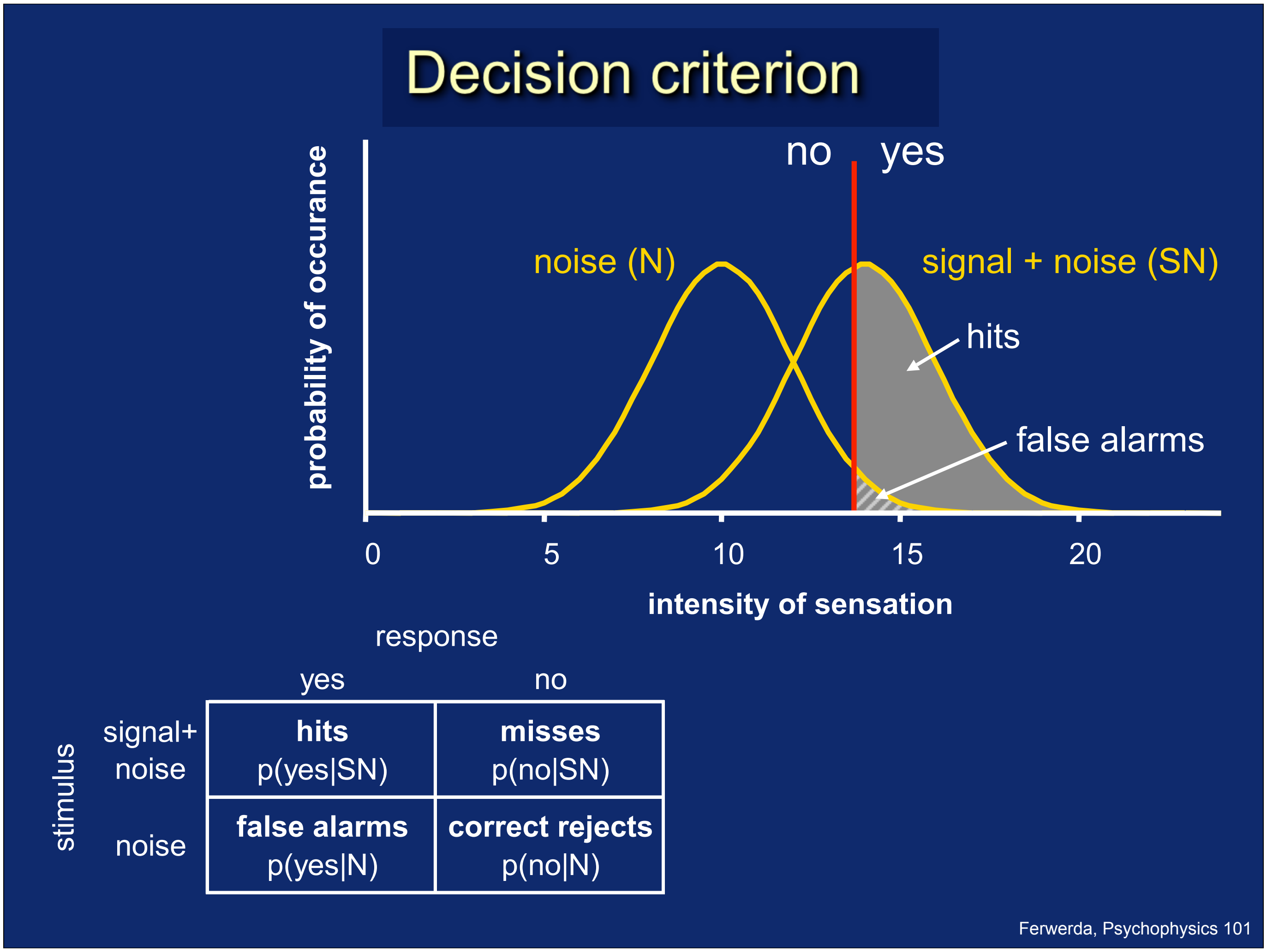
- if you’re the radar operator in this situation, you’ve got a problem
- all you experience on each sweep is a some intensity of sensation
- but as the graph shows, because of the overlap in the distributions, intensity could be caused by a real signal or could be due to noise alone
- no way to objectively distinguish between these two situations
- operator needs to make a DECISION that intensities greater than some level represent a true signal and values below are only noise
- outcome of this decision process is directly related to the threshold values we measure in psychophysical experiments
- to understand variation we see in threshold measures, need to look closer at operator’s decision-making processes

stimulus/response (SR) matrix

		response	
		yes	no
stimulus	signal+ noise	**hits** p(yes|SN)	**misses** p(no|SN)
	noise	**false alarms** p(yes|N)	**correct rejects** p(no|N)



- consider that each sweep of the beam is a trial in a threshold experiment
- on each trial there is a stimulus
  - signal embedded in noise
  - noise alone
- and on each trial observer has to give a response
  - yes – there was a signal
  - no – there wasn't a signal
- for each trial, four possible outcomes
  - signal, yes – hit
  - signal, no – miss
  - no signal, yes – false alarm
  - no signal, no – correct reject
- over many trials we can calculate the probabilities that the observer will respond in a particular way
  - notation is shorthand, e.g. hit
  - p(yes|SN) – probability of observer giving a "yes" response, given that a signal was actually present
- experimental results tabulated this way form stimulus/response (SR) matrix



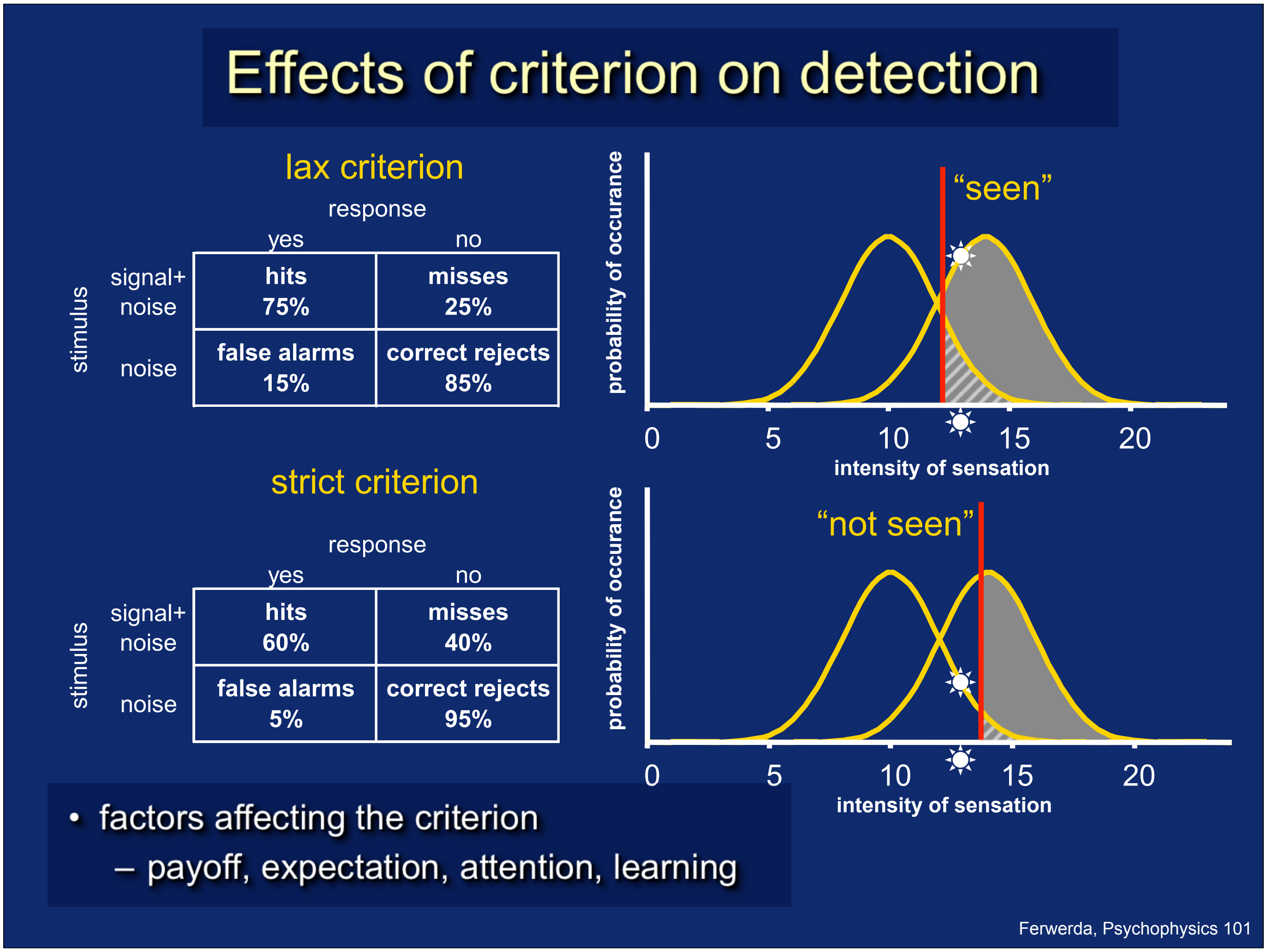
Decision criterion

<table>
<tr><td></td><td></td><td colspan="2">response</td></tr>
<tr><td></td><td></td><td>yes</td><td>no</td></tr>
<tr><td rowspan="2">stimulus</td><td>signal+ noise</td><td>hits
p(yes|SN)</td><td>misses
p(no|SN)</td></tr>
<tr><td>noise</td><td>false alarms
p(yes|N)</td><td>correct rejects
p(no|N)</td></tr>
</table>

Ferwerda, Psychophysics 101

- probabilities in the SR matrix correspond to areas under the noise and signal+noise distributions
- for example:
  - gray area under SN = hits
  - striped area under N = false alarms
  - unshaded areas correspond to misses and correct rejects respectively

- boundary of these areas define a location along the intensity axis
- location represents the observer’s criterion for deciding whether a particular intensity represents a signal or not
- notice:
  - responses above criterion are “yes’s” (seen)
  - responses below criterion are “no’s” (not seen)
  - yes/no responses independent of whether the signal was actually present
  - i.e misses and false alarms (observer’s performance not perfect)



Effects of criterion on detection

lax criterion

		response	
		yes	no
stimulus	signal+ noise	**hits 75%**	**misses 25%**
	noise	**false alarms 15%**	**correct rejects 85%**

strict criterion

		response	
		yes	no
stimulus	signal+ noise	**hits 60%**	**misses 40%**
	noise	**false alarms 5%**	**correct rejects 95%**

- factors affecting the criterion
  – payoff, expectation, attention, learning

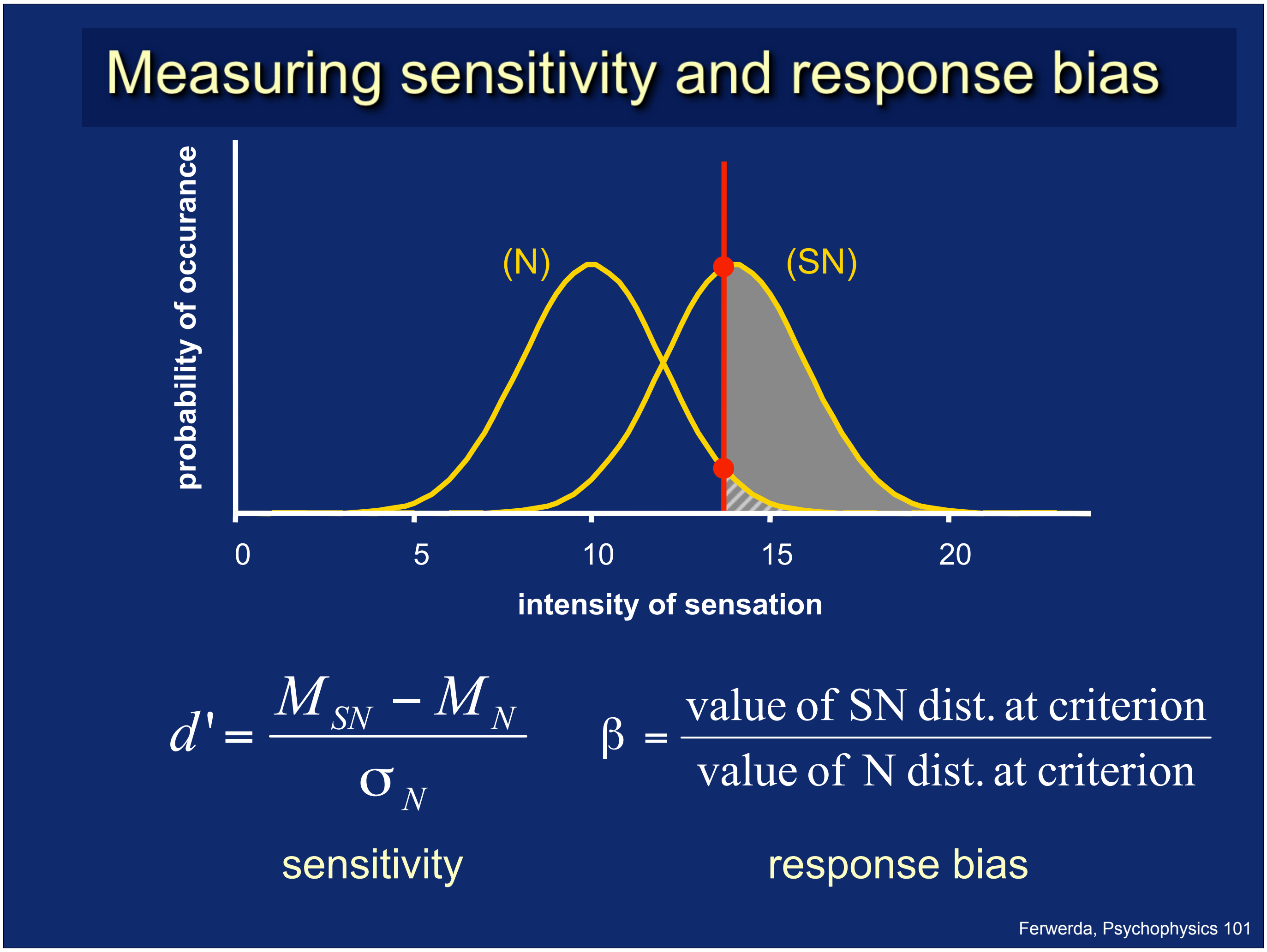
Ferwerda, Psychophysics 101

- so how does the observer’s decision criterion influence the thresholds we measure in psychophysical experiments
- imagine an experiment where we present a fixed intensity signal on each trial

- top figure, can see what happens when observer adopts a lax criterion for detection
  - willing to say “yes” when not sure, e.g. (going on alert in missle silo)
  - hit rate – high, misses – low, false alarms – relatively high
  - observer would report sensation intensity as “seen”
  - in experiment would be above threshold

- in contrast, bottom figures shows what happens when observer adopts a strict criterion
  - observer wants to be sure before saying “yes”, e.g. (firing missles!)
  - hit rate lower, misses higher, false alarms lower
  - observer would report given sensation intensity as “not seen”
  - in experiment would be below threshold

  - many psychological and situational factors affecting observer’s decision criterion
  - cost/payoff, expectation/frequency,, attention/vigilance/fatigue, learning/experience



Measuring sensitivity and response bias

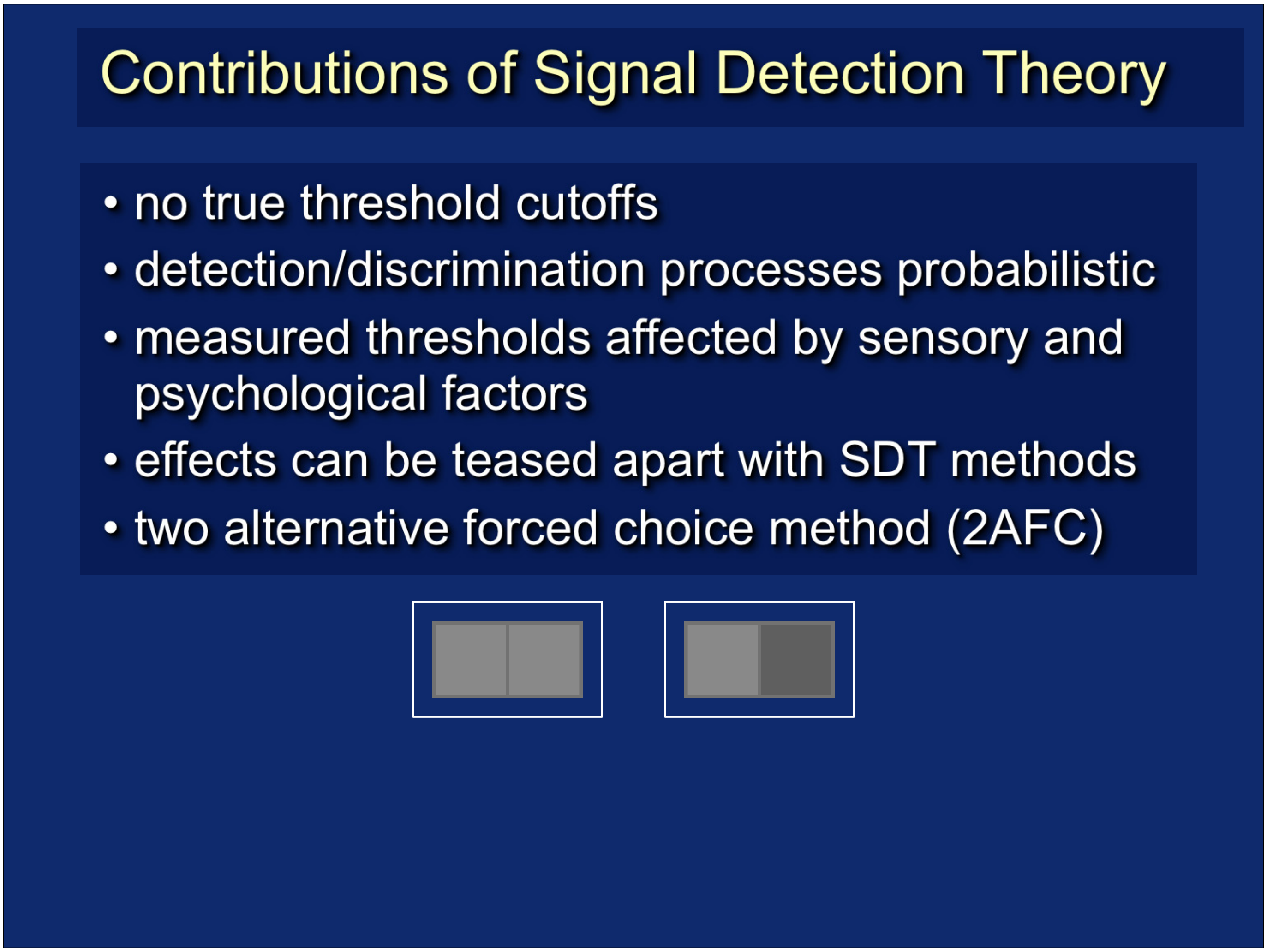
$$d' = \frac{M_{SN} - M_N}{\sigma_N} \qquad \beta = \frac{\text{value of SN dist. at criterion}}{\text{value of N dist. at criterion}}$$

sensitivity

response bias

Ferwerda, Psychophysics 101

- important contribution of SDT…
- SDT allows us to derive two measures that separate the observer’s underlying sensitivity from their decision making processes
- called d’ and beta
- beta = measure of the observers response bias – propensity to say yes/no when uncertain = ratio of values of SN and N at criterion
- d’ = measure of observer’s “true” sensitivity independent of any decision making factors = distance between means of N / SN distributions scaled by variance in distributions
  - = related to amount of overlap of distributions, potential for uncertainty whether sensation is signal or noise
- measures allow us to tease apart the sensory and non-sensory factors that affect thresholds measured in experiments



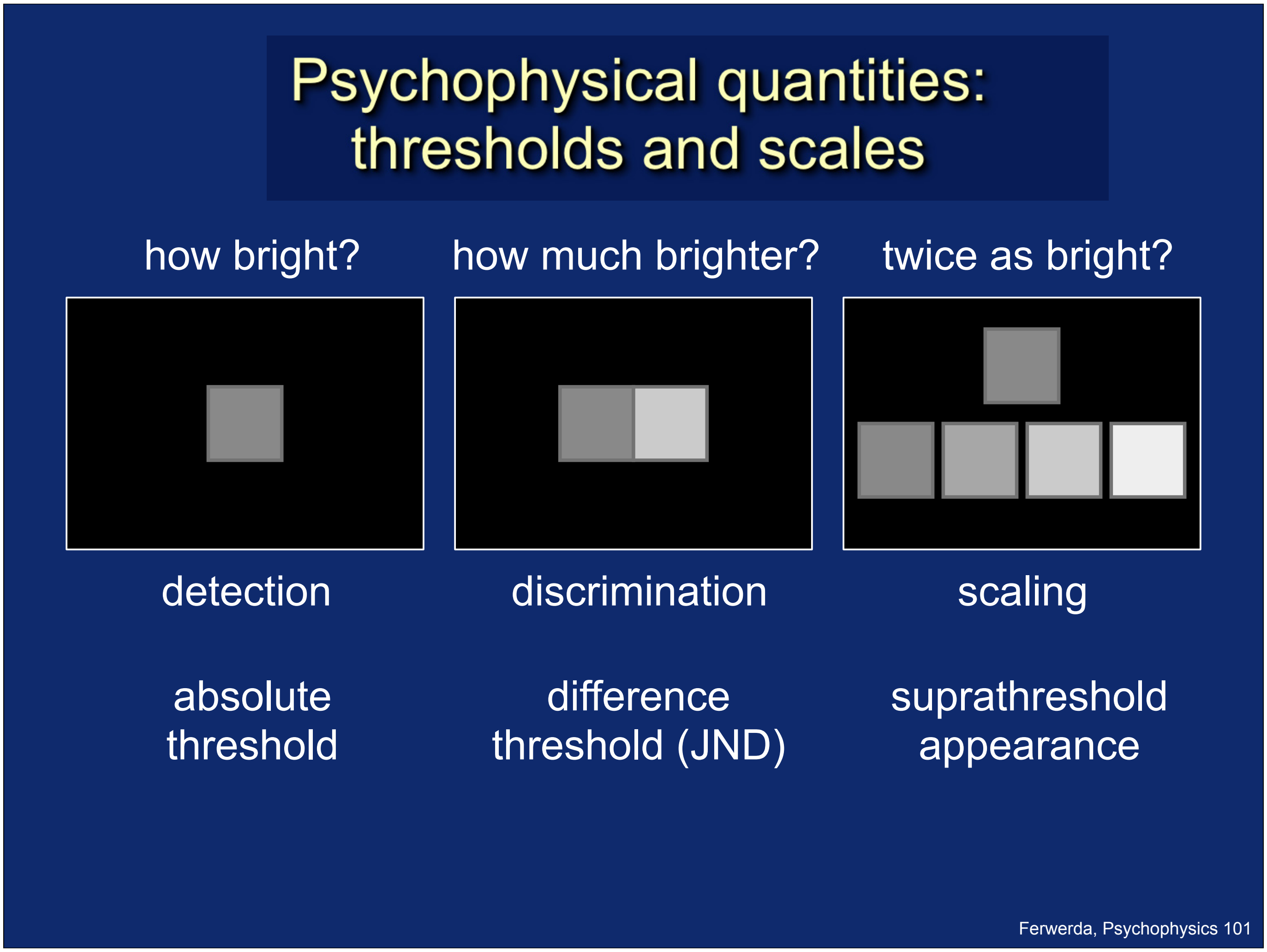
Contributions of Signal Detection Theory

- no true threshold cutoffs
- detection/discrimination processes probabilistic
- measured thresholds affected by sensory and psychological factors
- effects can be teased apart with SDT methods
- two alternative forced choice method (2AFC)

- got into this topic trying to understand variation seen in threshold measures
- SDT makes several contributions to understanding

- no true/sharp threshold cutoffs between seen/not seen
- detection and discrimination processes near threshold are probabilistic due to noise (external/internal)
- thresholds measured in experiments are a product of both sensory and non-sensory psychological factors (decision making)
- effects of these factors can be teased apart using SDT methods (d’, beta)

- influence of SDT has led to the development of important variation on classical threshold method …
- known as two-alternative forced choice (2AFC) method
- figure illustrates method for a difference threshold experiment
- on each trial two pairs of stimuli presented
- observer has to indicate which pair contained the difference
- allows direct measurement of guessing under conditions of uncertainty, allows more direct measures of underlying sensitivity

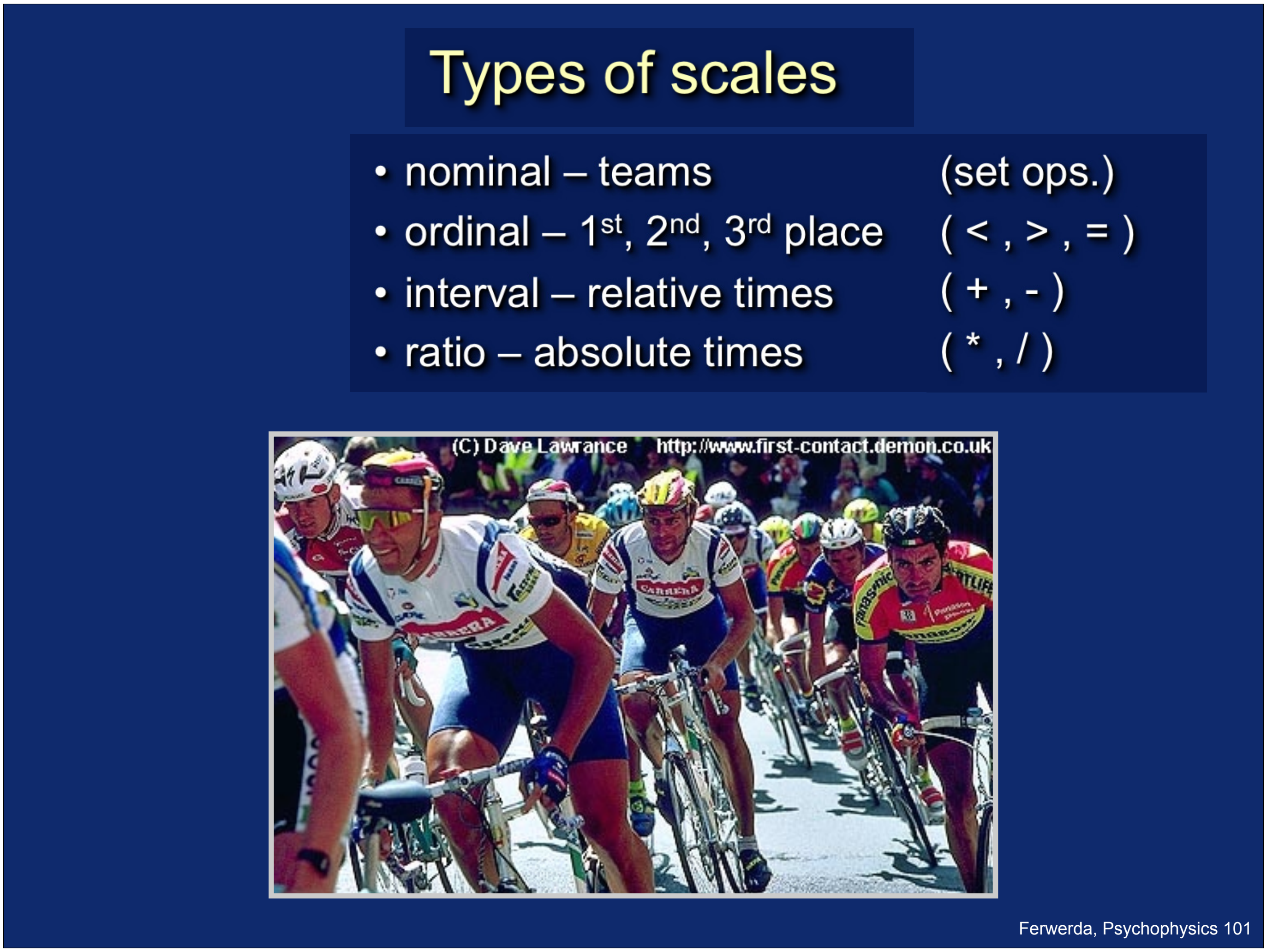


Psychophysical quantities: thresholds and scales

how bright?	how much brighter?	twice as bright?
detection	discrimination	scaling
absolute threshold	difference threshold (JND)	suprathreshold appearance

- so this brings us to the end of the section on threshold methods
- take home messages:
  - two kinds of thresholds: absolute/detection, difference/discrimination
  - variety of experimental methods for measuring thresholds:
    - Fechner
    - adjustment, limits, constant stimuli
    - SDT

- next section will be talking about psychophysical scaling
- whereas threshold methods concerned with measuring the limits of perception
- scaling deals with everything else
- goal is to quantify the relationships between the physical properties of objects and their perceptual appearances
- typical scaling issue: which is twice as bright?
- e.g. might use results of scaling experiments to build a scale relating physical radiance and perceived brightness

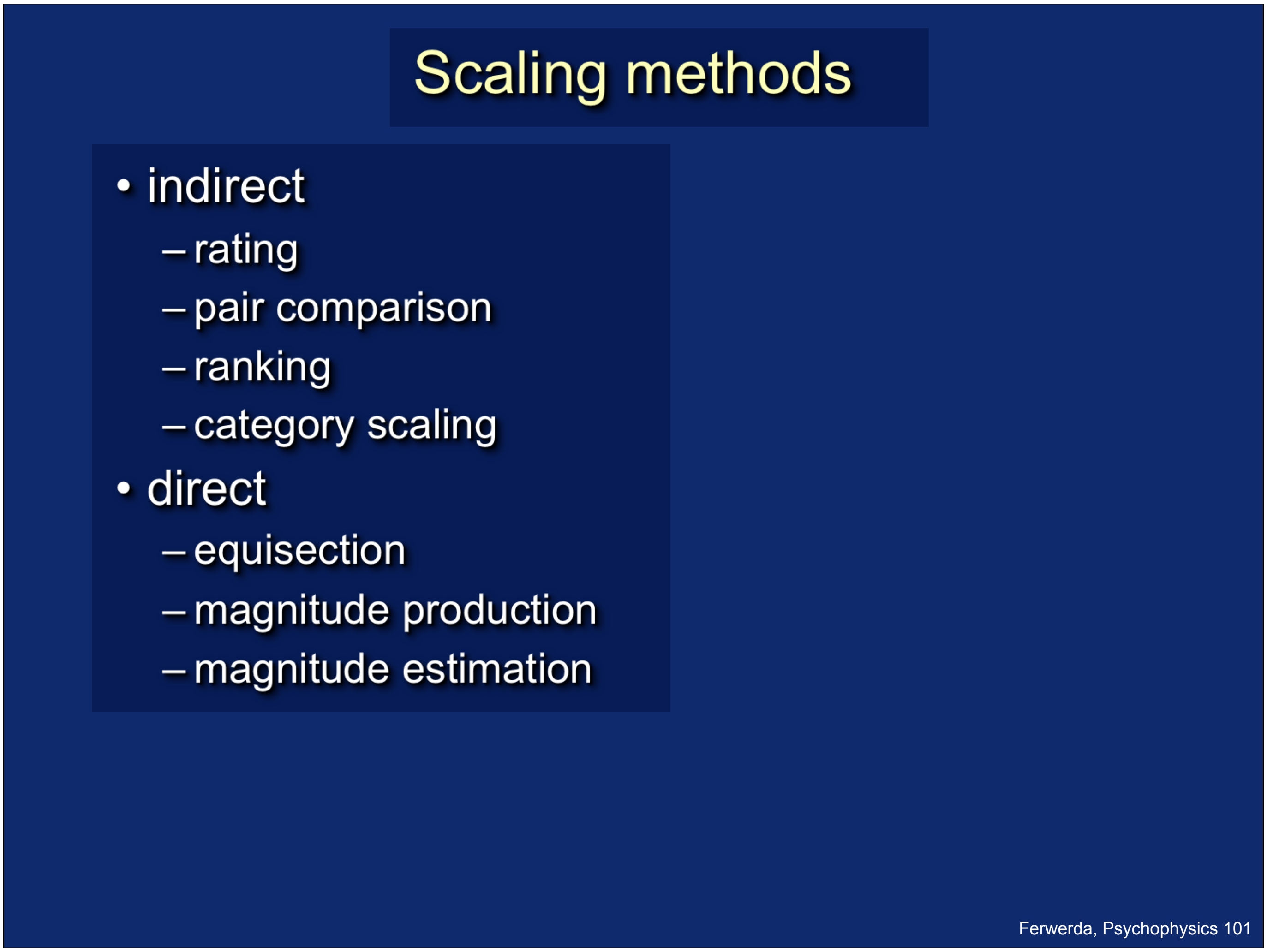


Types of scales

- nominal – teams (set ops.)
- ordinal – 1st, 2nd, 3rd place (< , > , =)
- interval – relative times (+ , -)
- ratio – absolute times (* , /)

(C) Dave Lawrance http://www.first-contact.demon.co.uk

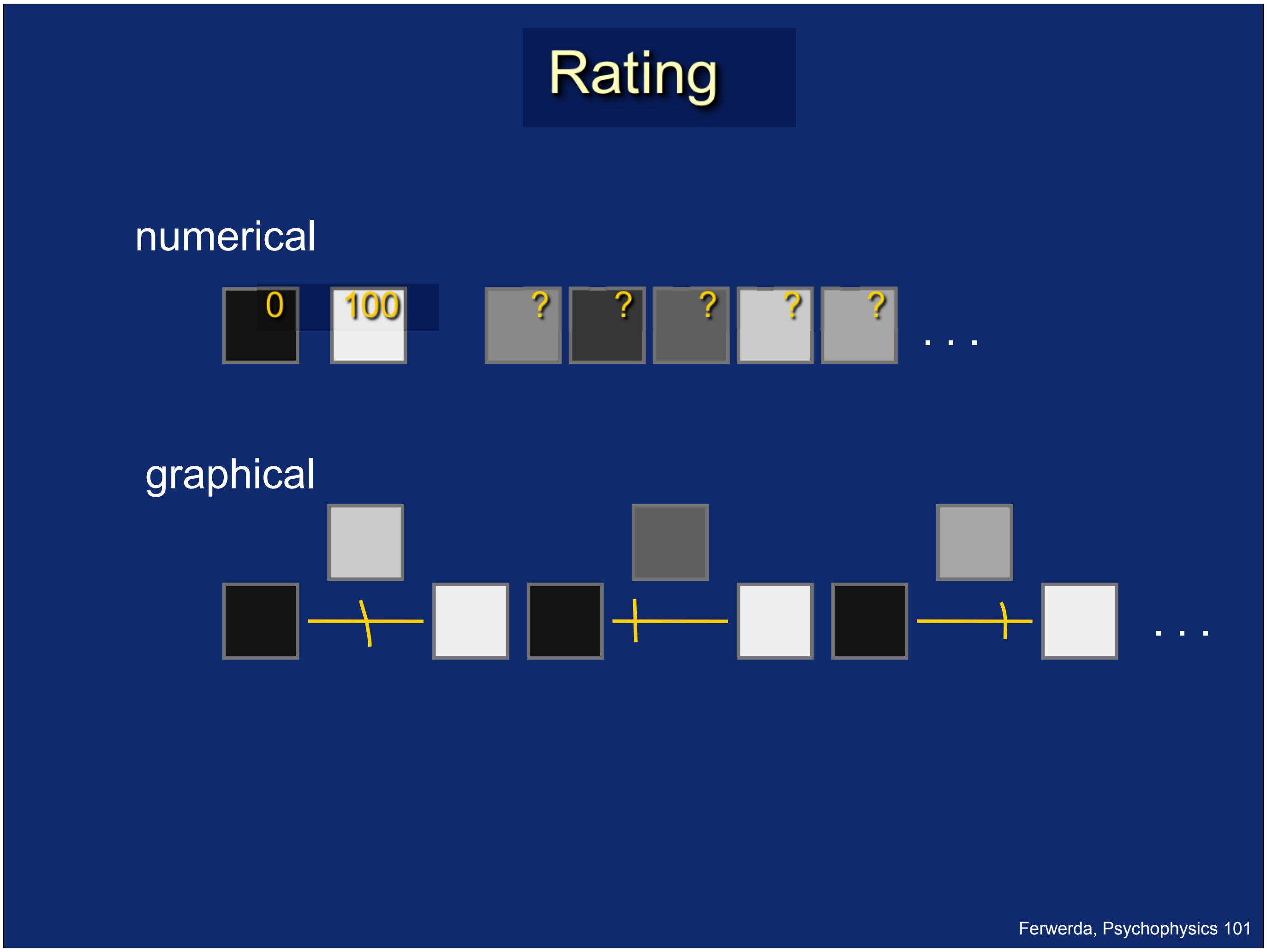
- first thing to understand:
  - many different types of scales we can use to measure properties of objects
  - different scales have different mathematical properties
- bike race as illustration
  - NOMINAL: simplest scale – nominal = naming
  - organize racers by teams (france, italy, us)
  - allows group/categorize like objects
  - doesn’t allow us to say anything quantitative about relations between groups
  - ORDINAL: allows ordering of objects w.r.t some criterion
  - e.g. who finished 1st, 2nd,3rd, …
  - ordinal scales allow objects to be ranked but numbers are only labels, cannot be used in calculations
    - e.g. 2nd place finisher did not necessarily take twice as long as 1st place
  - INTERVAL: to measure distances (space or time) need to construct interval scale
  - e.g. relative finishing times (2nd finished 5 mins after 1st, 3rd finished 4 mins after 2nd)
  - allows us to tell how far apart things are but not whether the differences are small or large in any absolute sense
  - e.g 5 mins difference between 1st and 2nd in Tour de France might be small, but in a 10 mile race would be large
  - RATIO: to understand magnitude of differences need a ratio scale
  - e.g absolute elapsed times of racers
  - comparing absolute times allows us to say
    - 2nd rider took 10% longer than 1st, 3rd place took twice as long as second
  - important feature of ratio scales: have a zero point
  - allows calculation of magnitudes



Scaling methods

- indirect
  - rating
  - pair comparison
  - ranking
  - category scaling
- direct
  - equisection
  - magnitude production
  - magnitude estimation

- wide variety of methods have been developed to derive psychophysical scales
- can be divided into two main categories: indirect and direct
- indirect:
  - subjects are asked to make simple judgments about object properties (eg. grouping/sorting/ordering)
  - numerical scale values are derived using statistics on these judgments
- direct:
  - subjects directly assign numerical values to their perceptions, or adjust stimuli to stand in some mathematical relation
  - numerical scales are a direct byproduct of the experiment, minimal statistical analysis needed
- division relates to an ongoing debate in psychology about whether people use numbers linearly
  - e.g. jelly bean problem: good at estimating if numbers (or range) are small, but accuracy decreases for larger magnitudes

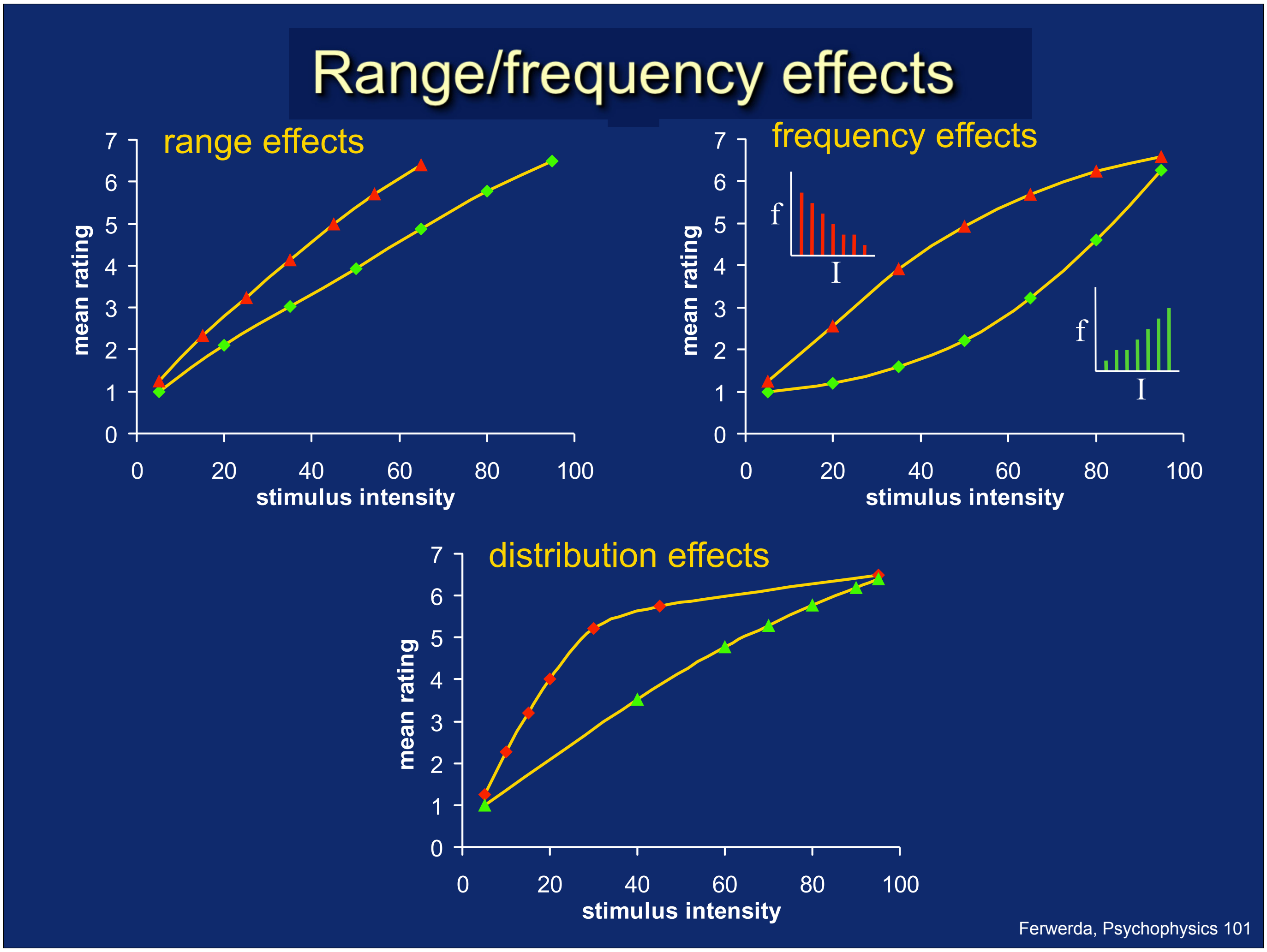


Rating

numerical

graphical

- first is rating
- two forms: numerical, graphical
  - numerical:
    - subject shown examples of endpoints of scale
    - asked: “on a scale of 0 to 100 tell me how bright this patch is”
- graphical:
  - subject shown endpoint examples connected by a line segment
  - asked to make a mark representing where the test patch would fall on an imaginary scale connecting the endpoints
  - numerical ratings derived by measuring distance from one endpoint
  - leverages fact that over short distances real length and perceived length are directly related



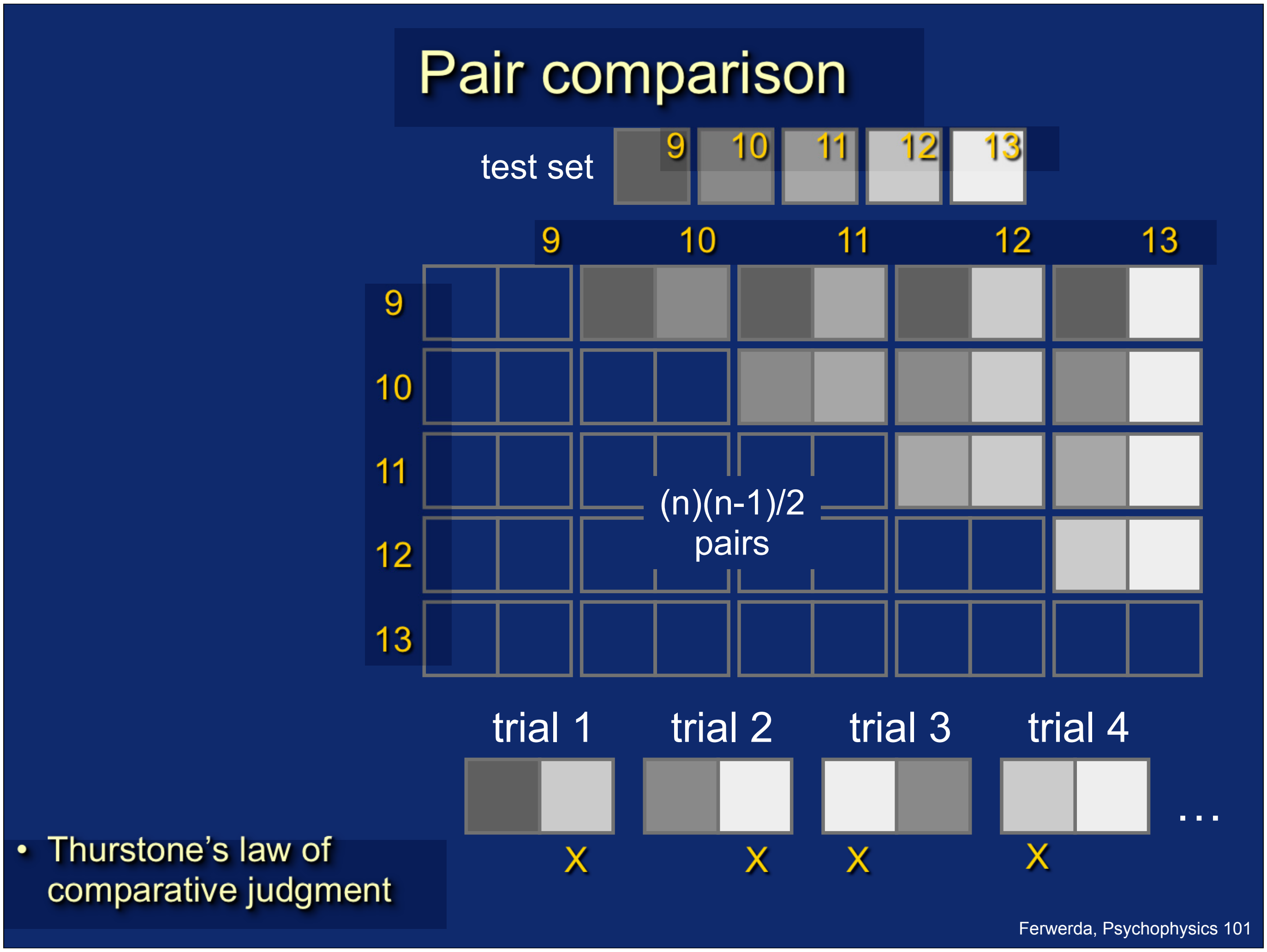
Range/frequency effects

Ferwerda, Psychophysics 101

- rating methods are widely used, generally well understood, liked by subjects
- but subject to a number of problems (range/frequency effects) that make them less than reliable for building psychophysical scales, effects illustrated in the graphs
- each data point shows the mean rating value given to some physical stimulus in a scaling experiment, the red and green curves show the scales obtained for different experiments

- first graph shows what are called “range effects”
  - range of intensities used in the green experiment was larger than in red experiment (0-100 vs. 0-60), differences in range lead to scales with different slopes
  - results of experiment affected by choice of range of test stimuli
  - shouldn’t be the case if the method was accurately measuring the true relationship between physical stimulus intensity and its perceptual appearance
- second graph shows “frequency effects”
  - red and green insets show how frequently each of the test stimuli was shown to the subject over the course of the experiment, red experiment: low intensities shown more frequently than hi, green: vice versa
  - differences in the shapes of the curves show that the scales are affected by how often subjects see the different stimuli
  - again this shouldn’t be the case if the method is accurately measuring the psychophysical relationship
- bottom graph shows “distribution effects”
  - in red experiment stimuli are mostly clustered near the low intensity range with a few high intensity values, green experiment opposite: mostly high intensity test stimuli with few low values
  - unequal distribution of test stimuli along the range leads to different scale estimates
  - again this shouldn’t be the case if the rating method were reliable

- doesn’t mean that rating methods are all bad there are methods for controlling these effects
- need to be aware of these problems if you decide to use rating methods in your experiments



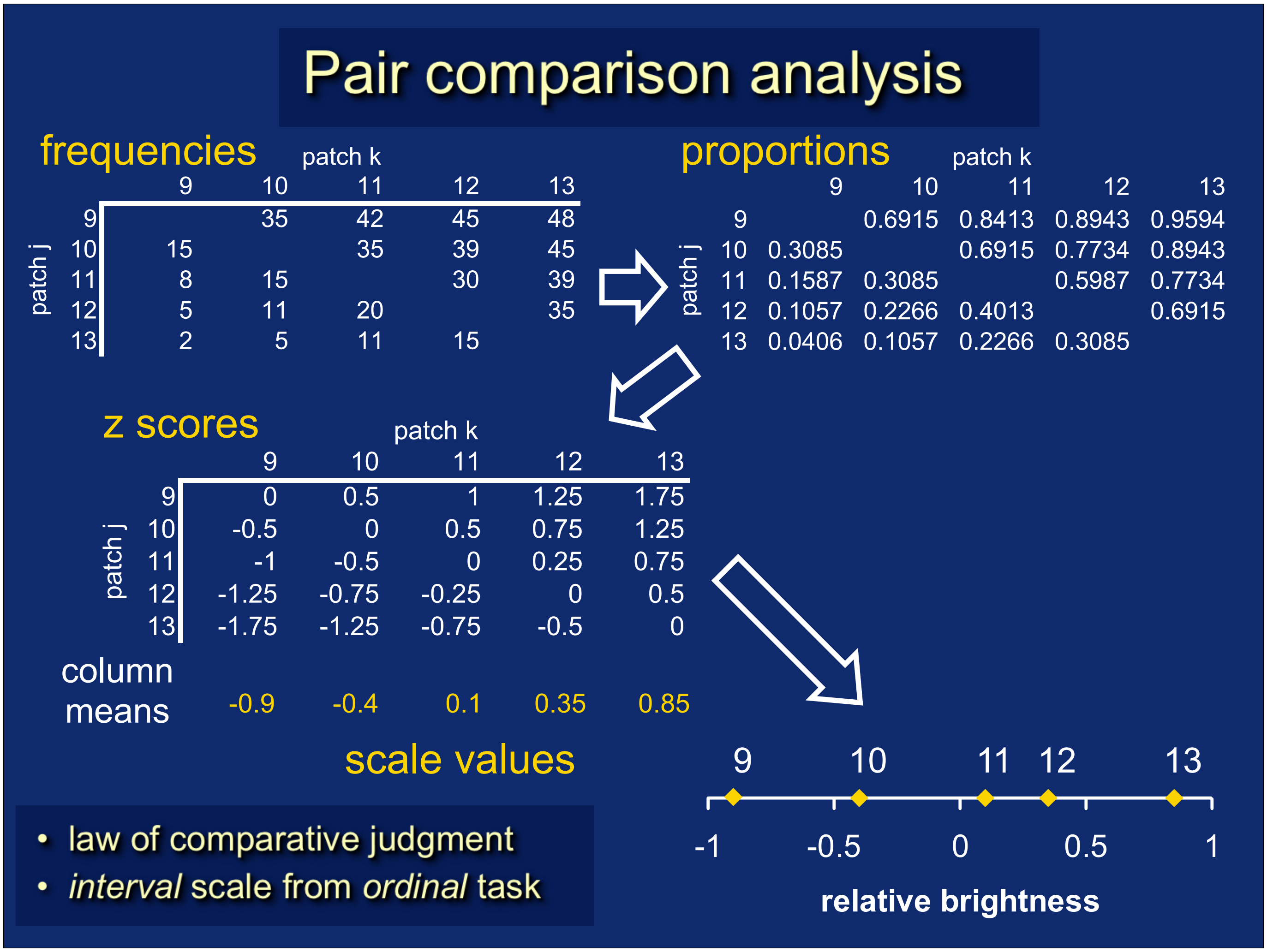
Pair comparison

- Thurstone’s law of comparative judgment

Ferwerda, Psychophysics 101

- second indirect scaling method is pair comparison
- as name implies stimuli from the test set are organized into pairs as shown in the table
- on each trial one pair is presented and subject is asked (for example) to say which one is brighter
- two important issues about choosing the test stimuli
  - number of pairs scales as (n)(n-1)/2 so if set is large number of trials can get large too
  - values need to be close enough together so subject is confused/makes mistakes (as in trial 4) since statistics for deriving scale values depend on some pair differences being sub-threshold

- to derive a psychophysical scale from the results of this experiment, can use an analysis method called the “law of comparative judgment”
- this and the “law of categorical judgment” were developed by Thurstone in the 20’s as part of his general theory of scaling
- Thurstone is the father of modern scaling methods, similar in stature to Fechner



Pair comparison analysis

frequencies (patch j × patch k)

patch j \ patch k	9	10	11	12	13
9		35	42	45	48
10	15		35	39	45
11	8	15		30	39
12	5	11	20		35
13	2	5	11	15	

proportions

patch j \ patch k	9	10	11	12	13
9		0.6915	0.8413	0.8943	0.9594
10	0.3085		0.6915	0.7734	0.8943
11	0.1587	0.3085		0.5987	0.7734
12	0.1057	0.2266	0.4013		0.6915
13	0.0406	0.1057	0.2266	0.3085	

z scores

patch j \ patch k	9	10	11	12	13
9	0	0.5	1	1.25	1.75
10	-0.5	0	0.5	0.75	1.25
11	-1	-0.5	0	0.25	0.75
12	-1.25	-0.75	-0.25	0	0.5
13	-1.75	-1.25	-0.75	-0.5	0
column means	-0.9	-0.4	0.1	0.35	0.85

scale values

- law of comparative judgment
- *interval* scale from *ordinal* task

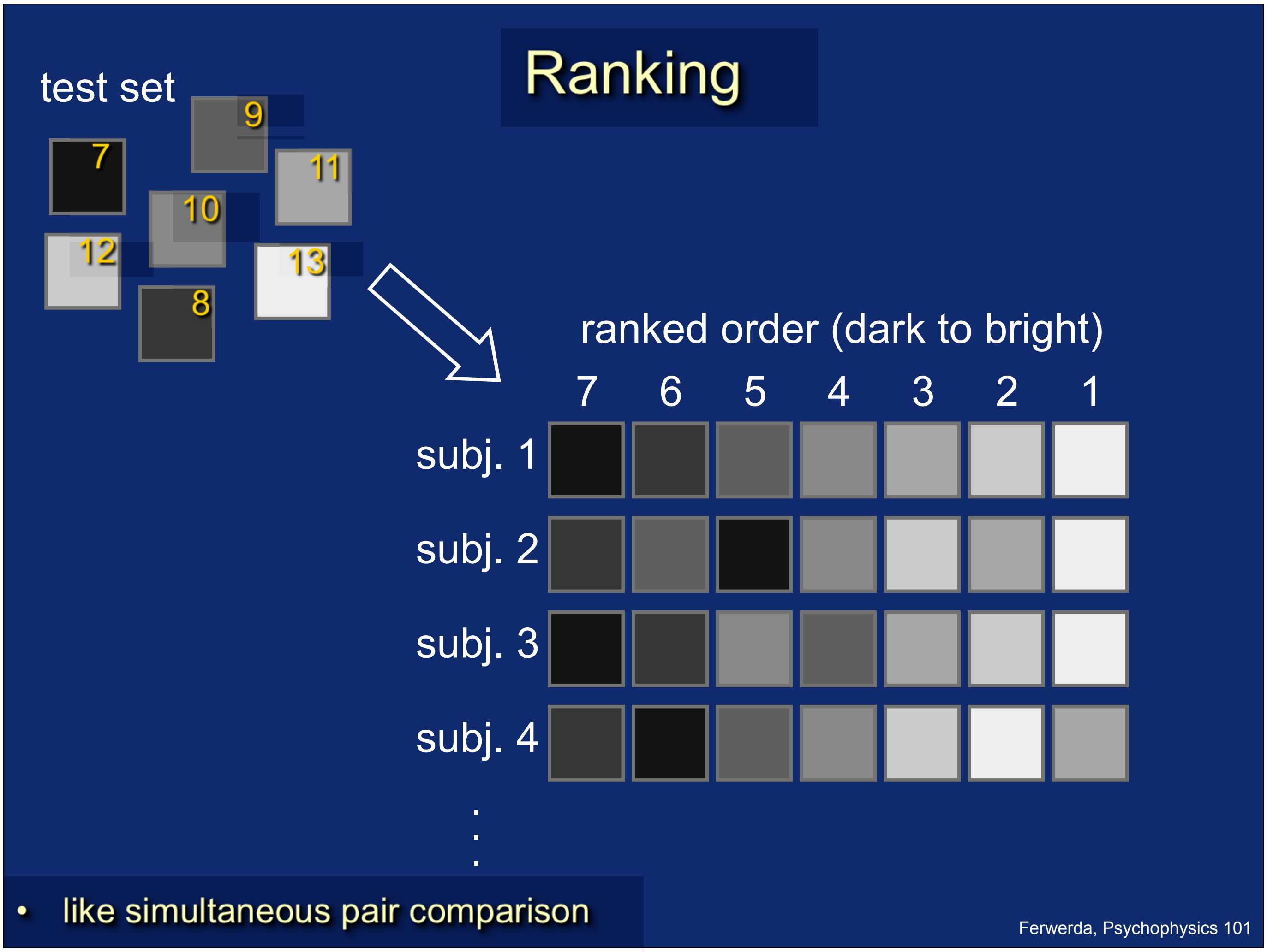
- whole books written about Thurstonian scaling and the law of comparative judgment (some listed in bibliography)
- will skip over theory and just describe the mechanics of the analysis

- first tabulate the data into the frequency matrix
- each cell indicates the number of times one patch was judged brighter than the other
- e.g. (10 vs. 9, 35 vs. 15)
- next convert these frequencies into proportions by dividing by the number of trials for each pair (in this case 50)
- next, assuming that the data are normally distributed, can convert the proportions to z-scores
- column means of the z-scores correspond to the values of the test stimuli on an interval scale

- notice that this is an interval scale
  - distances between stimuli indicate how similar the test patches are perceived to be in brightness
    - e.g. 11 and 12 seen as very close
  - BUT…scale is in terms of relative brightness (relative to mean of set)
    - no true zero, can’t say 13 is perceived to be twice as bright as 10 (need a ratio scale)

- take home message: by using this analysis (law of comparative judgment, normally distributed data) we’re able to derive an interval scale of brightness from ordinal judgments (pair comparisons)

Ranking



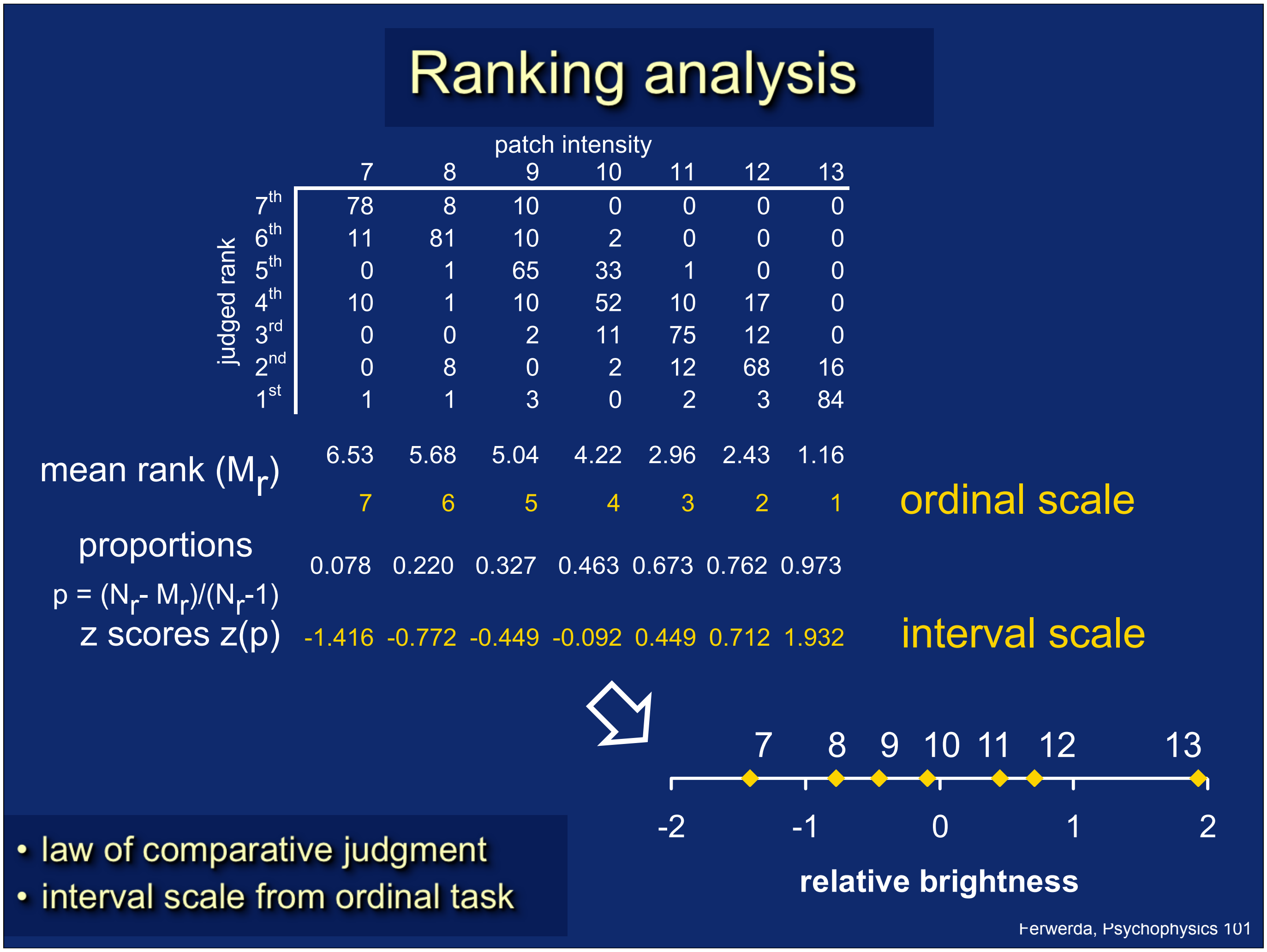
- like simultaneous pair comparison

Ferwerda, Psychophysics 101

- pair comparison is a good method, but can often lead to many trials, boring for subjects, problematic effects of fatigue
- another indirect scaling method that can provide similar results is ranking

- subject is shown entire test set in scrambled order
- asked to place the patches in order (for example from dark to bright)
- subjects generally like ranking experiments (puzzle-like)
- happy subjects provide good data

- figure on lower right shows the results produced by four subjects
- notice that subjects are not complete accurate in their rankings (e.g. some reversals)
- also not complete agreement between subjects
- perceptual scale can be derived by drawing on Thurstone’s law of comparative judgment



Ranking analysis

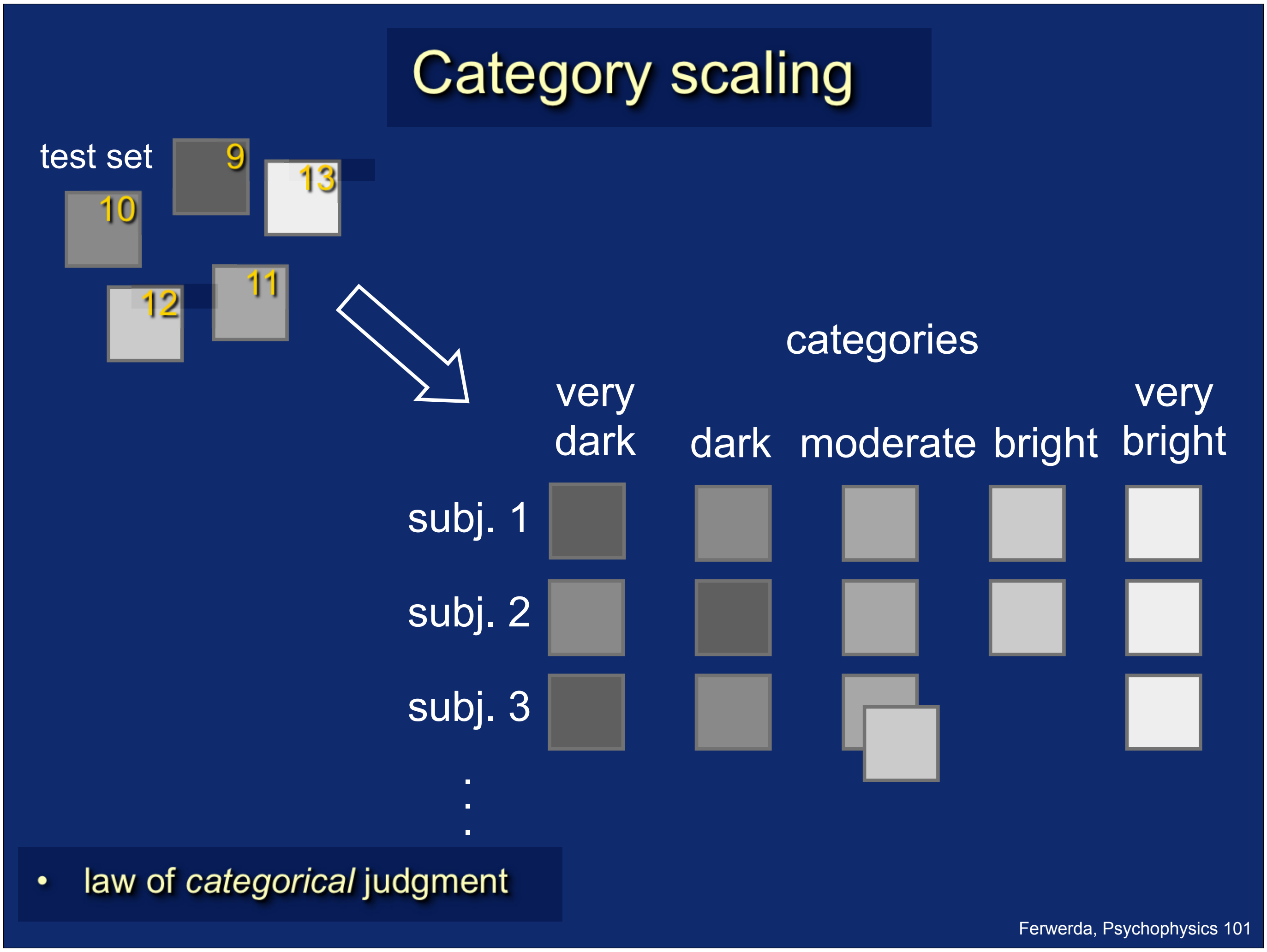
	patch intensity						
judged rank	7	8	9	10	11	12	13
7th	78	8	10	0	0	0	0
6th	11	81	10	2	0	0	0
5th	0	1	65	33	1	0	0
4th	10	1	10	52	10	17	0
3rd	0	0	2	11	75	12	0
2nd	0	8	0	2	12	68	16
1st	1	1	3	0	2	3	84

mean rank (M_r)	6.53	5.68	5.04	4.22	2.96	2.43	1.16	
	7	6	5	4	3	2	1	ordinal scale
proportions $p = (N_r - M_r)/(N_r - 1)$	0.078	0.220	0.327	0.463	0.673	0.762	0.973	
z scores z(p)	-1.416	-0.772	-0.449	-0.092	0.449	0.712	1.932	interval scale

- law of comparative judgment
- interval scale from ordinal task

Ferwerda, Psychophysics 101

- first step: tabulate the data showing how often each patch was placed in each rank position (data from 100 subjects)
- notice that least intense consistently ranked lowest, but some spread, similar for increasing intensities
- next: calculate the mean rank (7*78 + 6*11 + … /100)
- mean rank values can be used directly to derive an ordinal scale (no surprises, monotonic)
- but mean ranks and ordinal scale say nothing about the perceived differences between the patches (monotonic not same as linear)
- by again assuming a normal model and using Thurstone's law of comparative judgment we can derive an interval scale from the rank data
- first derive proportions (# of ranks – mean rank)/(# ranks – 1)
- then convert proportions to z scores
- z-scores correspond to values of patches on an interval scale of relative brightness
- notice again: equal physical spacing does not lead to equal perceived spacing

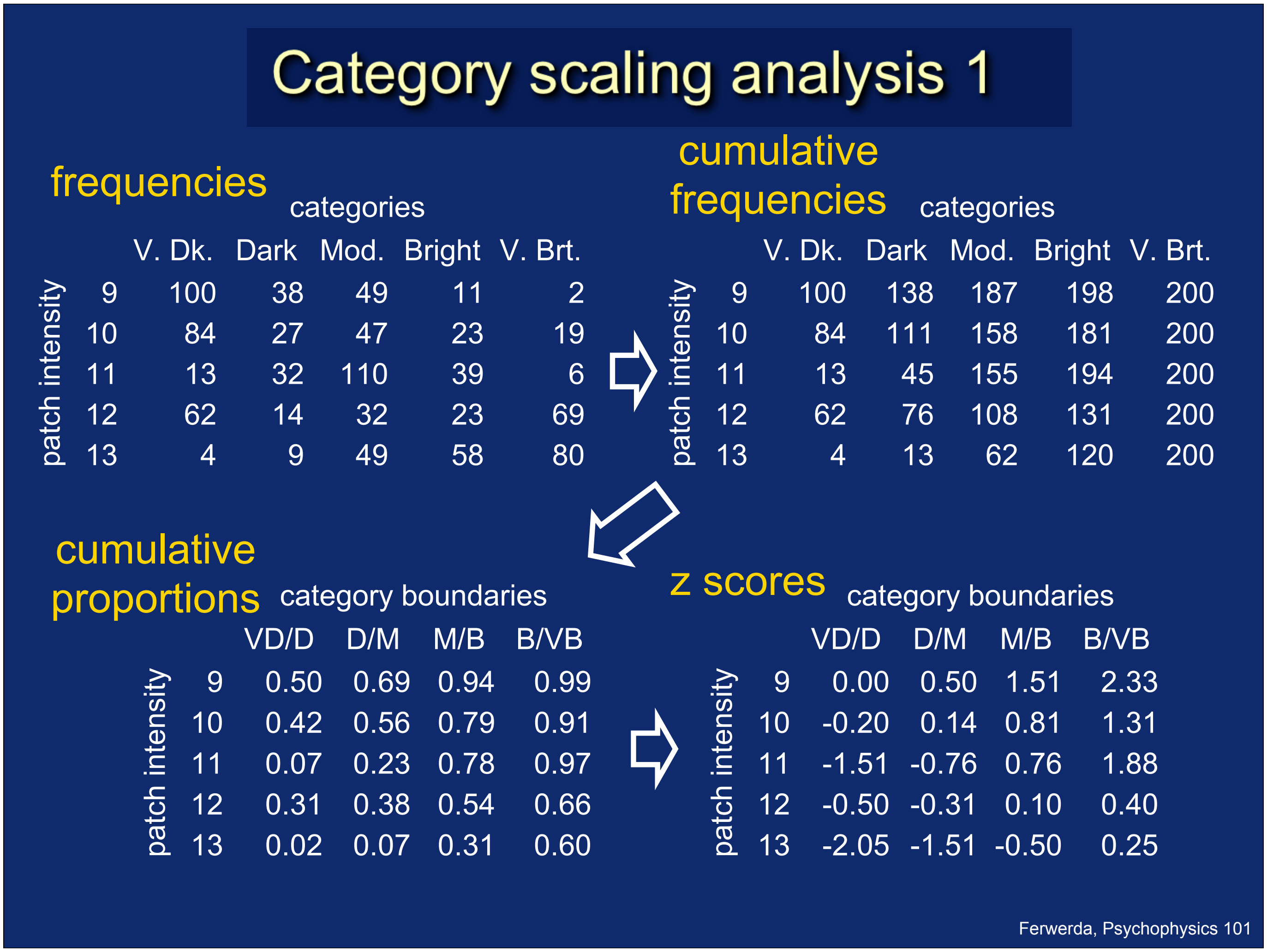


Category scaling

- law of *categorical* judgment

- last of indirect scaling methods we'll look at is category scaling
- experimental procedure is the same as categorical rating method discussed earlier
- subject shown members of test set and asked to assign them to labeled categories
- figure on lower right shows results from 3 subjects
- again notice confusions/reversals/non-agreement in categorization

- may have noticed that experimental procedure is the same as the categorical rating method I showed earlier that I said was problematic
- difference between categorical rating and category scaling:
  - in categorical rating categories are assumed to be linear and equally spaced (not always true)
  - in category scaling can apply Thurstone's law of *categorical* judgement is applied to find the true shape and spacing of the scale



Category scaling analysis 1

frequencies

	categories				
patch intensity	V. Dk.	Dark	Mod.	Bright	V. Brt.
9	100	38	49	11	2
10	84	27	47	23	19
11	13	32	110	39	6
12	62	14	32	23	69
13	4	9	49	58	80

cumulative frequencies

	categories				
patch intensity	V. Dk.	Dark	Mod.	Bright	V. Brt.
9	100	138	187	198	200
10	84	111	158	181	200
11	13	45	155	194	200
12	62	76	108	131	200
13	4	13	62	120	200

cumulative proportions

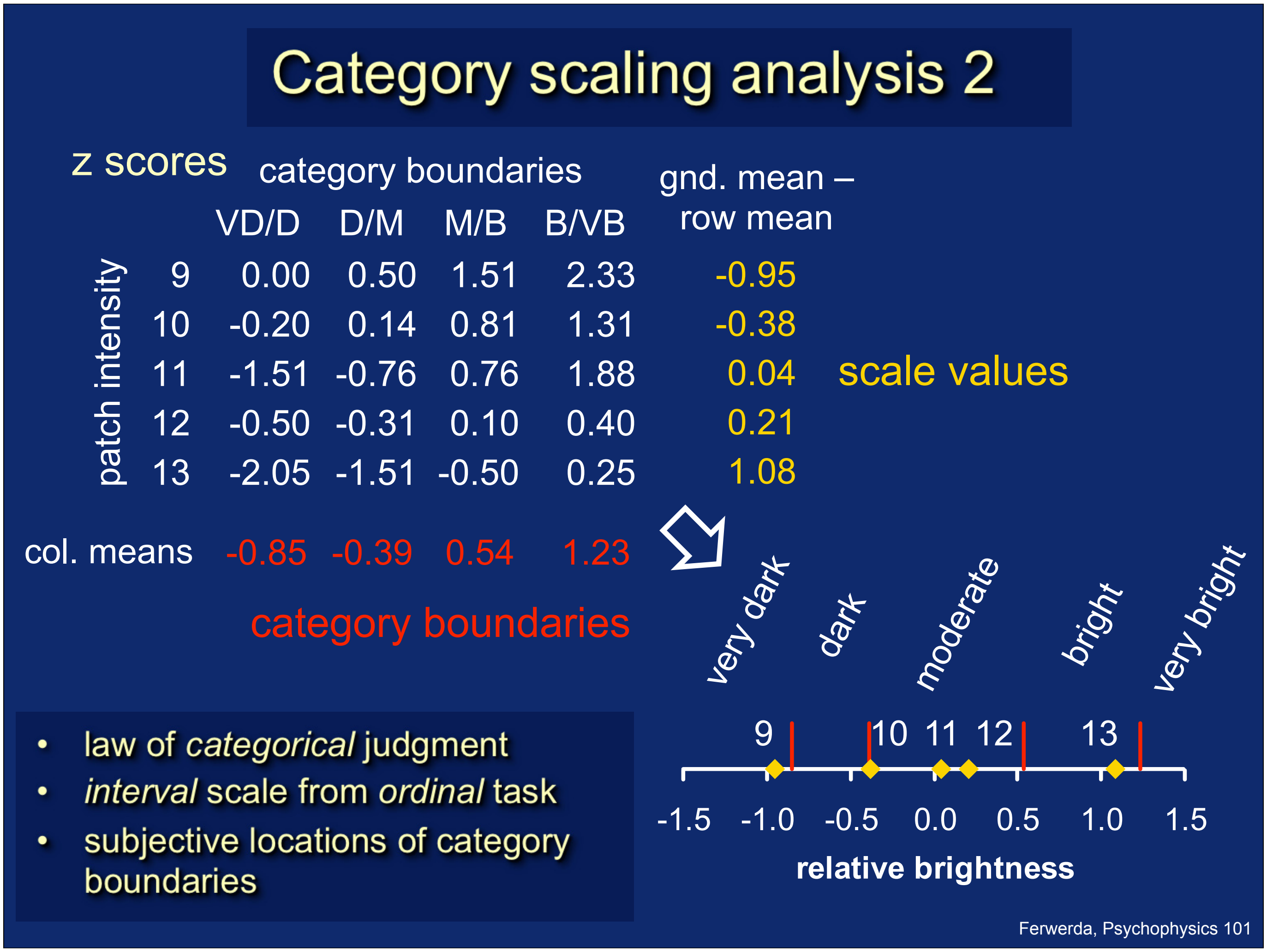
	category boundaries			
patch intensity	VD/D	D/M	M/B	B/VB
9	0.50	0.69	0.94	0.99
10	0.42	0.56	0.79	0.91
11	0.07	0.23	0.78	0.97
12	0.31	0.38	0.54	0.66
13	0.02	0.07	0.31	0.60

z scores

	category boundaries			
patch intensity	VD/D	D/M	M/B	B/VB
9	0.00	0.50	1.51	2.33
10	-0.20	0.14	0.81	1.31
11	-1.51	-0.76	0.76	1.88
12	-0.50	-0.31	0.10	0.40
13	-2.05	-1.51	-0.50	0.25

Ferwerda, Psychophysics 101

- category scaling analysis:
- first step is to tabulate the data
  - each cell indicates # of times a patch was placed into each category (200 trials/patch)
  - notice low intensity placed into dark categories, but some spread
- next calculate cumulative frequencies for each patch
  - e.g. patch 9 judged dark on 100 trials, dark or very dark on 138 trials, …
- next use cumulative frequencies to calculate cumulative proportions and z-scores that represent the boundaries between categories
  - e.g. value in cell 1,1, represents an estimate of the boundary between the “very dark and “dark” categories



Category scaling analysis 2

z scores

patch intensity

	category boundaries				
	VD/D	D/M	M/B	B/VB	gnd. mean – row mean
9	0.00	0.50	1.51	2.33	-0.95
10	-0.20	0.14	0.81	1.31	-0.38
11	-1.51	-0.76	0.76	1.88	0.04
12	-0.50	-0.31	0.10	0.40	0.21
13	-2.05	-1.51	-0.50	0.25	1.08
col. means	-0.85	-0.39	0.54	1.23	

scale values

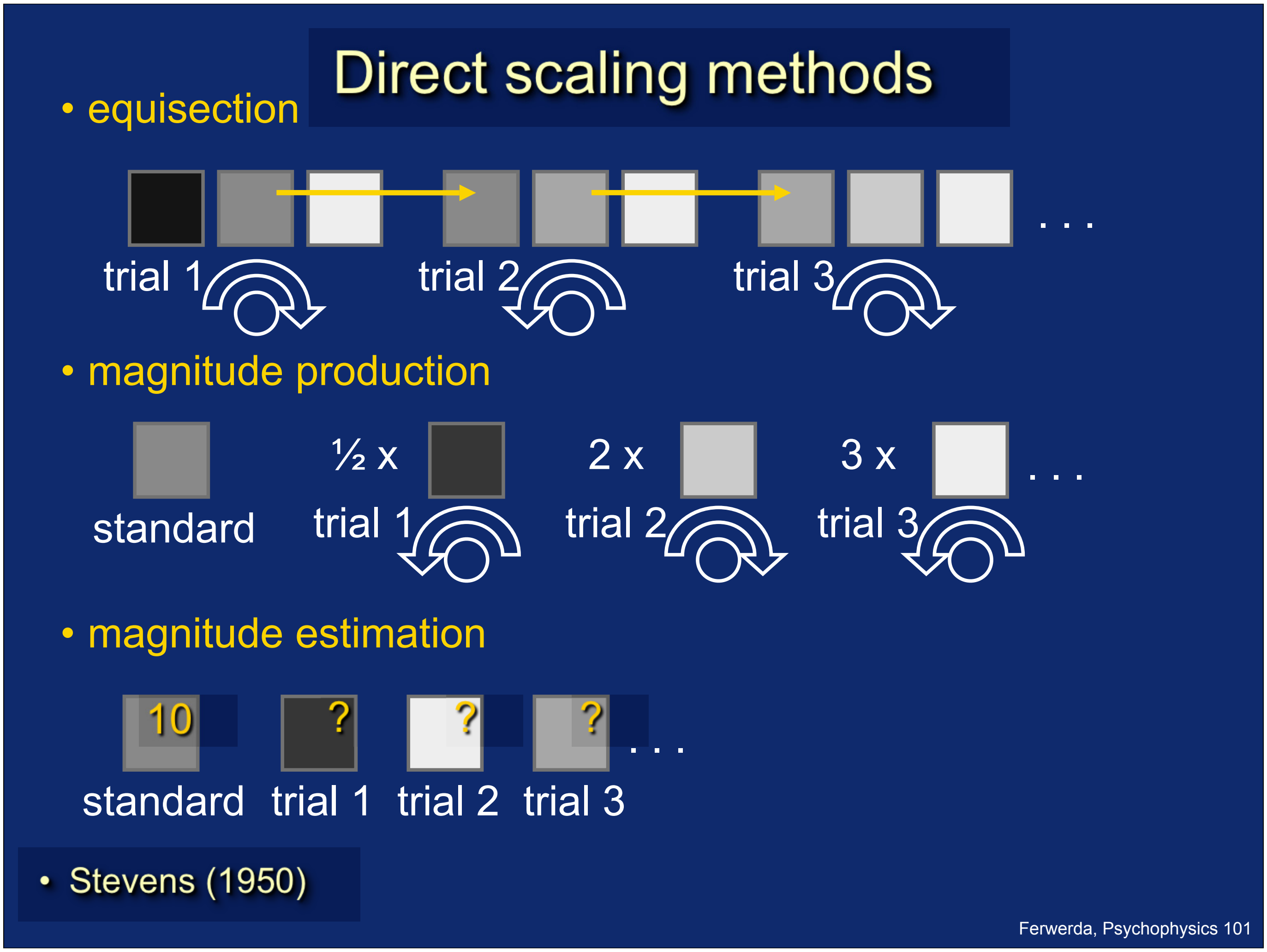
category boundaries

- law of *categorical* judgment
- *interval* scale from *ordinal* task
- subjective locations of category boundaries

Ferwerda, Psychophysics 101

- finally, using the z-scores and the applying the law of categorical judgment
  - can calculate the values of the patches on an interval scale of relative brightness (shown in orange)
  - can also calculate the apparent locations of the boundaries between the different categories used in the experiment (shown in red)
  - notice categories not equally spaced or equally wide
    - e.g. patch values 10,11,12 all perceived as “moderately bright”, none as “very bright”

- law of comparative judgment allows us to test the assumptions behind categorization methods and derive more meaningful scales



Direct scaling methods

- equisection

- magnitude production

- magnitude estimation

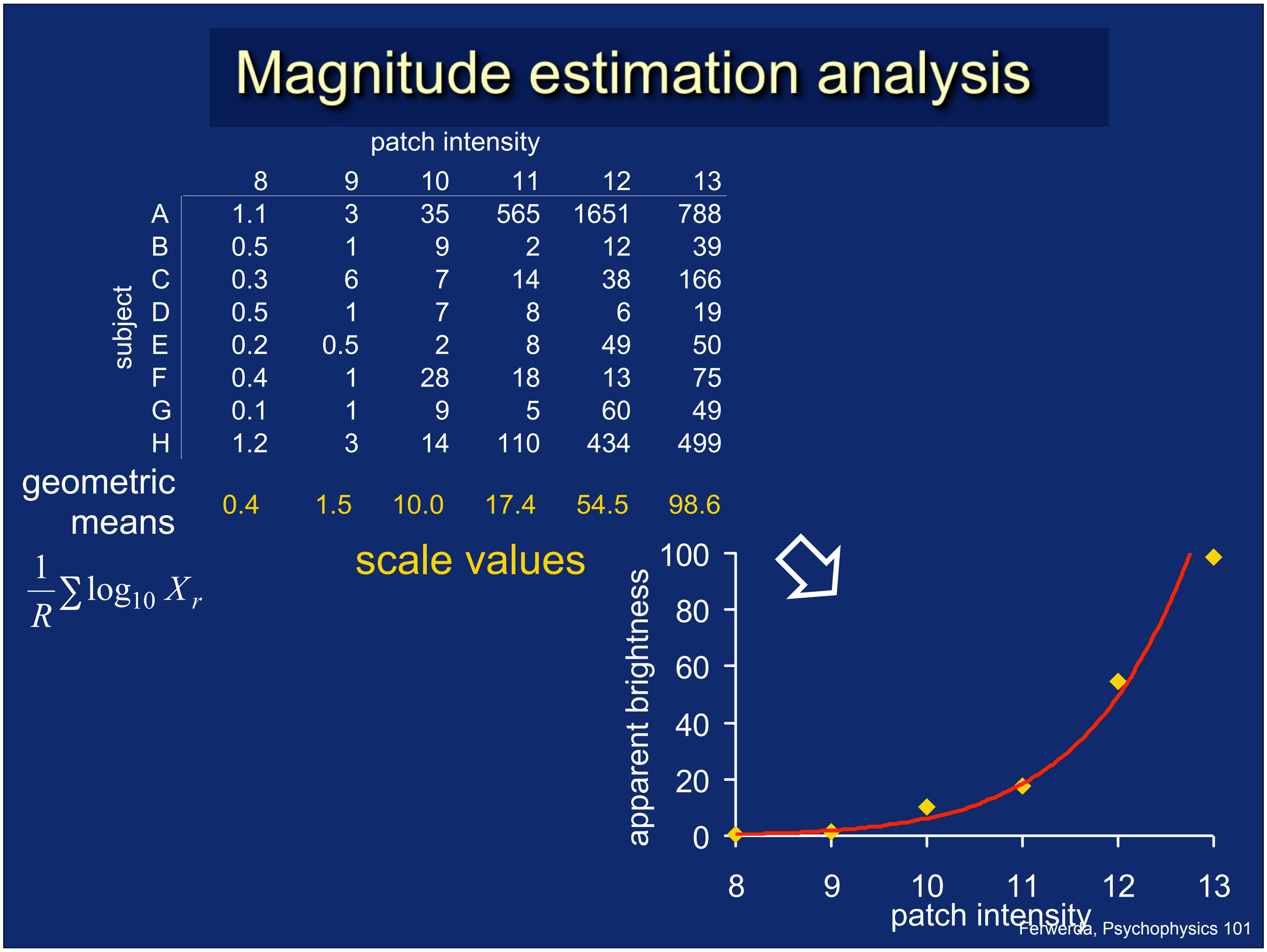
- Stevens (1950)

- indirect methods all based on having subjects make simple judgments of object ordinal properties (categorizing, ranking, rating), interval scales derived through statistical analysis (Thurstone's laws)
- in contrast, in direct methods subjects directly assign numbers to what they perceive or adjust stimuli to stand in particular numerical relationships
- direct methods are associated with S.S. Stevens who introduced/championed them in the '50's
- controversial because they rest on assumptions about how people use numbers, but beneficial because they can be used to develop ratio scales (most powerful)
- three direct methods:

- EQUISECTION: method of adjustment-like procedure
- test patch is bounded on either side by standards, subject is asked to adjust the brightness of the test patch until it falls halfway between the standards
- in second trial, test patch becomes the lower standard, and so on …
- method used to develop the munsell color space

- MAGNITUDE PRODUCTION: also an adjustment procedure
- subject shown a standard, then asked to adjust a test patch until it's half as bright, twice as bright, etc…
- settings provide direct estimates of a ratio scale

- MAGNITUDE ESTIMATION: most widely known of direct methods: magnitude estimation, most promoted by Stevens,
  - subject shown a standard, told has a brightness of "10", then shown series of test patches, asked to assign numbers representing apparent brightness w.r.t standard
  - choice of range of numbers is up to the subject, but studies show that after some settling down, subjects are internally consistent



Magnitude estimation analysis

	patch intensity					
subject	8	9	10	11	12	13
A	1.1	3	35	565	1651	788
B	0.5	1	9	2	12	39
C	0.3	6	7	14	38	166
D	0.5	1	7	8	6	19
E	0.2	0.5	2	8	49	50
F	0.4	1	28	18	13	75
G	0.1	1	9	5	60	49
H	1.2	3	14	110	434	499
geometric means	0.4	1.5	10.0	17.4	54.5	98.6

$$\frac{1}{R}\sum \log_{10} X_r$$

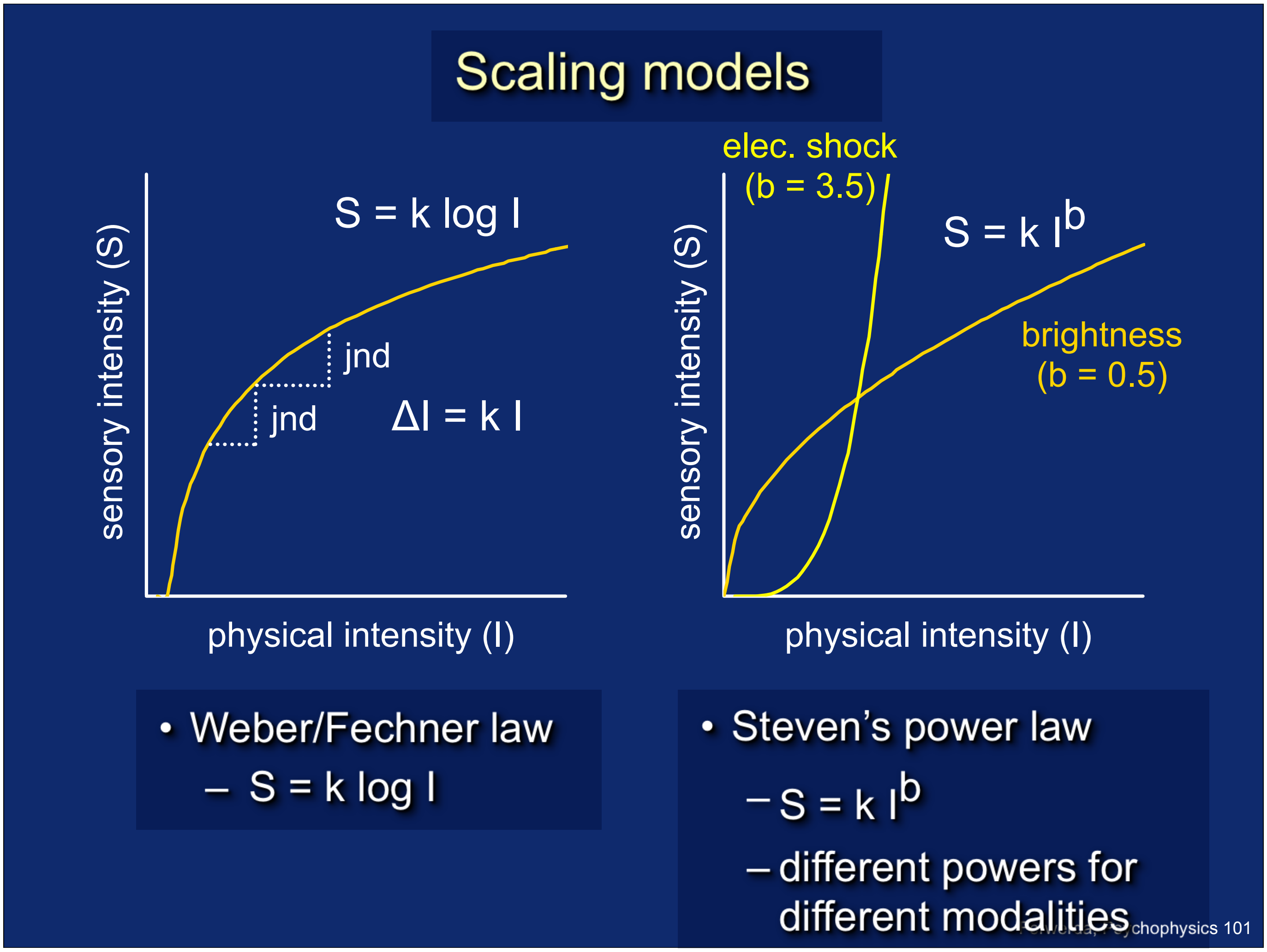
scale values

Ferwerda, Psychophysics 101

- analysis of the magnitude estimation data is straightforward
- simply involves averaging the estimates given by different subjects
- table shows magnitude estimates assigned by different subjects to our standard test set

- notice use of different numerical ranges by subjects
- e.g. (subject A 1.1 to 788), (subject B 0.5 to 39)
- to compensate for these differences, scale values are calculated using the geometric mean (average of logarithms of measured values)

- graph on right shows scale derived using magnitude estimation (relating apparent brightness to patch intensity)



Scaling models

- Weber/Fechner law
  - S = k log I

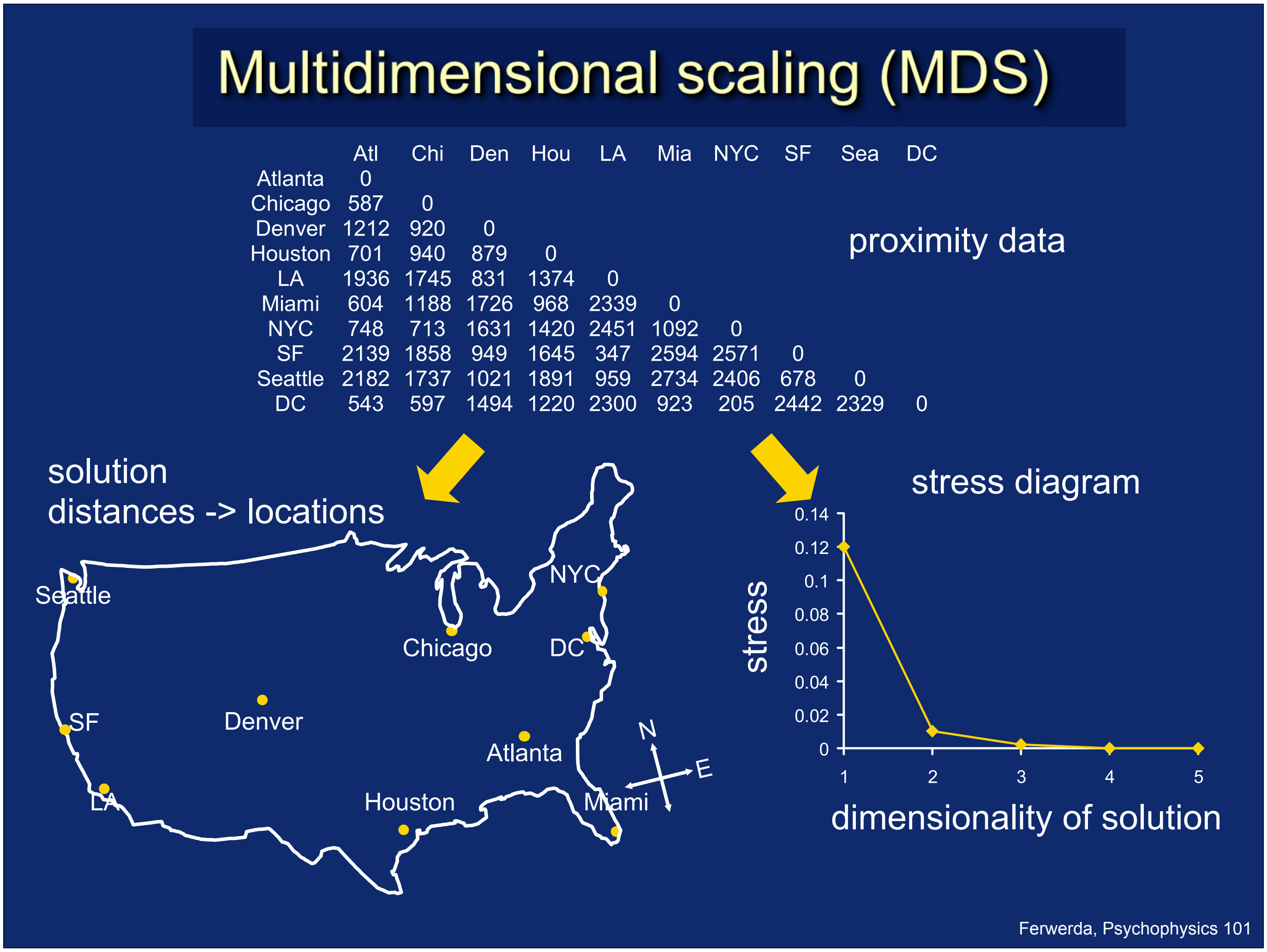
- Steven’s power law
  - $S = k\ I^b$
  - different powers for different modalities

- psychophysicists have tried to develop general mathematical models to describe psychophysical scales measured in experiments

- earliest model was Fechner’s law
  - based on Weber’s law – JNDs are proportional to intensity of standard
  - stacking up JNDs will produce a scale of perceptually uniform steps
  - Fechner’s scales are all of the form S = k log I (because of proportionality of JNDs)

- Steven’s measured psychophysical scales for many different sensory modalities using magnitude estimation techniques
  - (brightness, loudness, weight, heat/cold, electric shock)
  - found that his scales could all be described by a power function S = k I ^b
    - known as Stevens’ power law
    - different k’s and exponents for different sensory modalities
    - most are compressive functions
    - but some physically damaging ones like heat, electric shock have accelerating relations (protective mechanism?)

- ongoing debate over the years about whether Fechner’s or Steven’s law is correct scaling model
- central issue: validity of direct vs. indirect scaling methods



Multidimensional scaling (MDS)

	Atl	Chi	Den	Hou	LA	Mia	NYC	SF	Sea	DC
Atlanta	0									
Chicago	587	0								
Denver	1212	920	0							
Houston	701	940	879	0						
LA	1936	1745	831	1374	0					
Miami	604	1188	1726	968	2339	0				
NYC	748	713	1631	1420	2451	1092	0			
SF	2139	1858	949	1645	347	2594	2571	0		
Seattle	2182	1737	1021	1891	959	2734	2406	678	0	
DC	543	597	1494	1220	2300	923	205	2442	2329	0

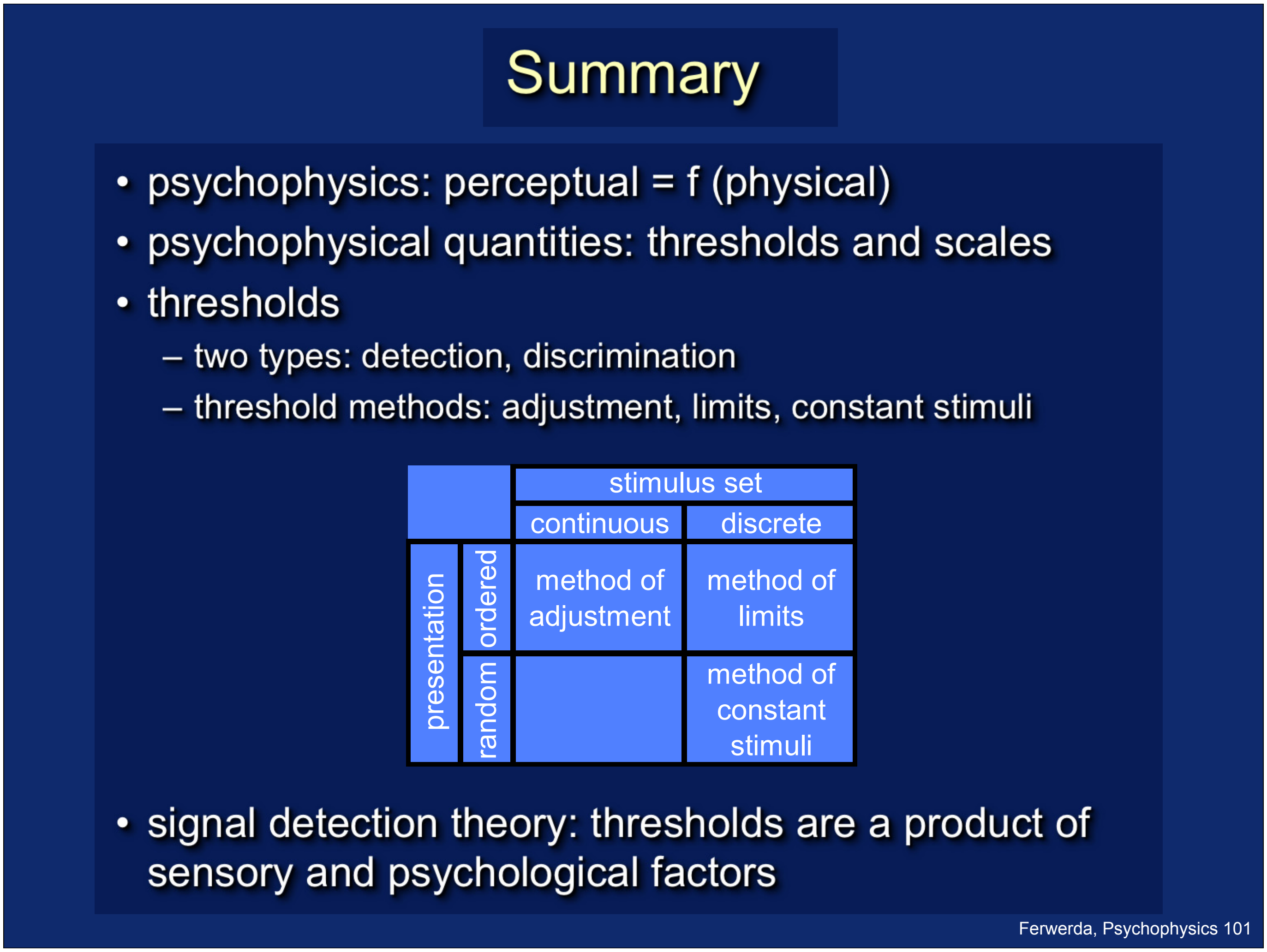
proximity data

solution
distances -> locations

stress diagram

Ferwerda, Psychophysics 101

- final scaling method: multidimensional scaling
- very useful tool for discovering psychophysical relations in complex stimuli (real objects, images)
- MDS is a method for finding the hidden structure in a dataset
- non-perceptual example
- given a table that specifies distances between objects in a dataset example: table shows distances between major US cities, table gives proximities but not spatial relations
- MDS will attempt to reconstruct a space that best describes it's overall structure
- picture on lower left shows the result produced by an MDS algorithm which has placed all the cities in their proper locations on the US map
- the graph on the right shows what is known as the stress diagram which indicates the goodness of fit of the MDS solution as dimensions are added to the solution. Notice the large drop in stress as we move from a 1-dimensional to a 2-dimensional solution because of the better fit to the data
- 3d is only slightly better (curvature of earth), so for most purposes 2d can be regarded as good solution
- classical MDS solutions only specific up to rotations, reflections, necessary element of interpretation in understanding MDS results
- MDS is useful for trying to understand the psychophysical dimensions of complex stimuli of the kind we often encounter in graphics



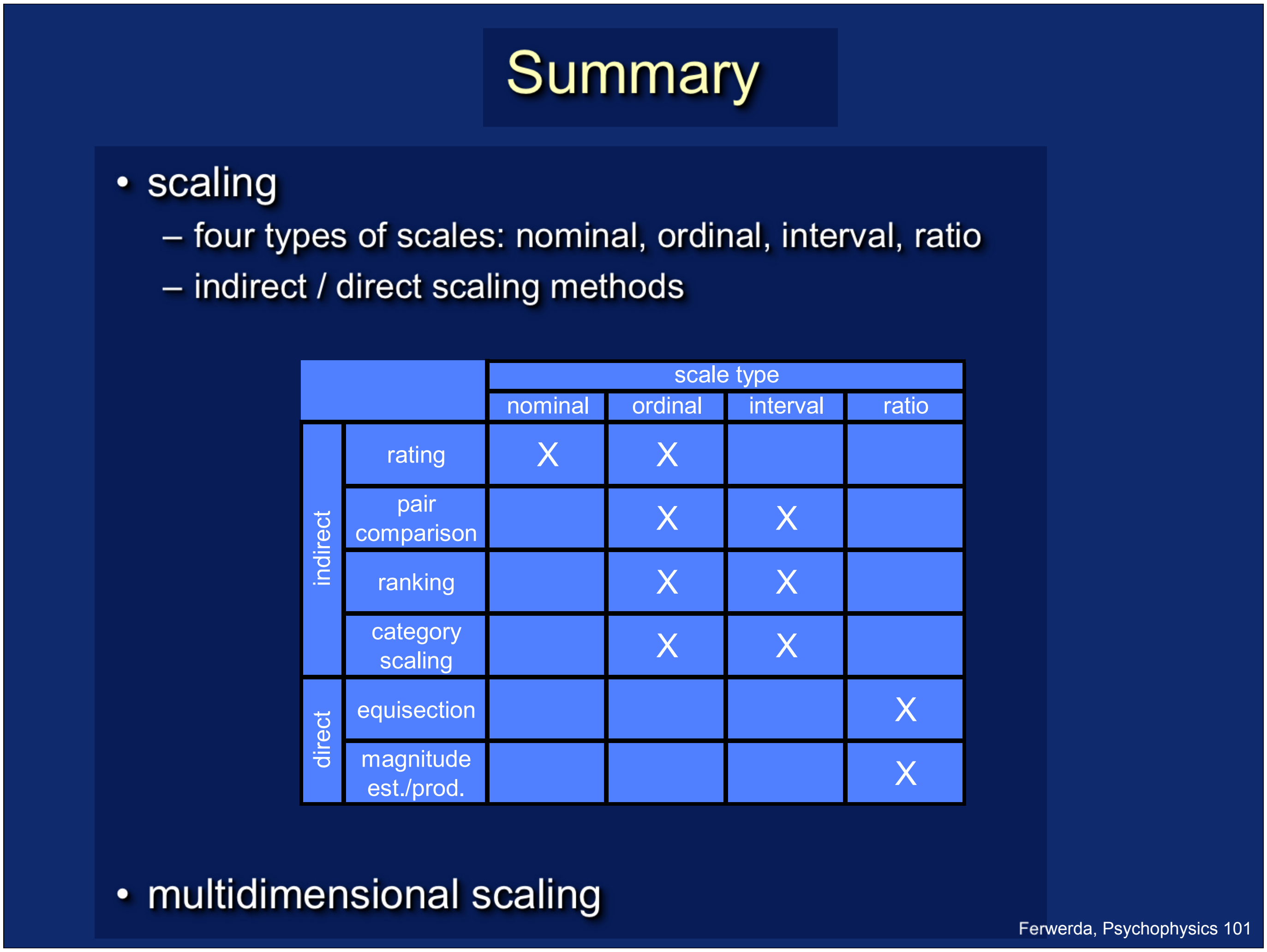
Summary

- psychophysics: perceptual = f (physical)
- psychophysical quantities: thresholds and scales
- thresholds
  - two types: detection, discrimination
  - threshold methods: adjustment, limits, constant stimuli

		stimulus set	
		continuous	discrete
presentation	ordered	method of adjustment	method of limits
	random		method of constant stimuli

- signal detection theory: thresholds are a product of sensory and psychological factors

- goal of psychophysics is to quantify the relationships between the physical properties of objects and their perceptual appearances
- two psychophysical quantities: thresholds and scales
- two types of thresholds: detection and discrimination
  - absolute sensitivity vs. ability to detect differences in visible stimuli
  - experimental methods introduced by Fechner: adjustment, limits, constant stimuli
  - table shows they are all really variations on the same theme, differ on…
    - stimulus set: continuous or discrete
    - stimulus presentation: ordered or randomized
- signal detection theory
  - contribution: thresholds are a product of sensory and psychological factors (cost/benefit, expectations)
  - SDT provides tools to tease these apart



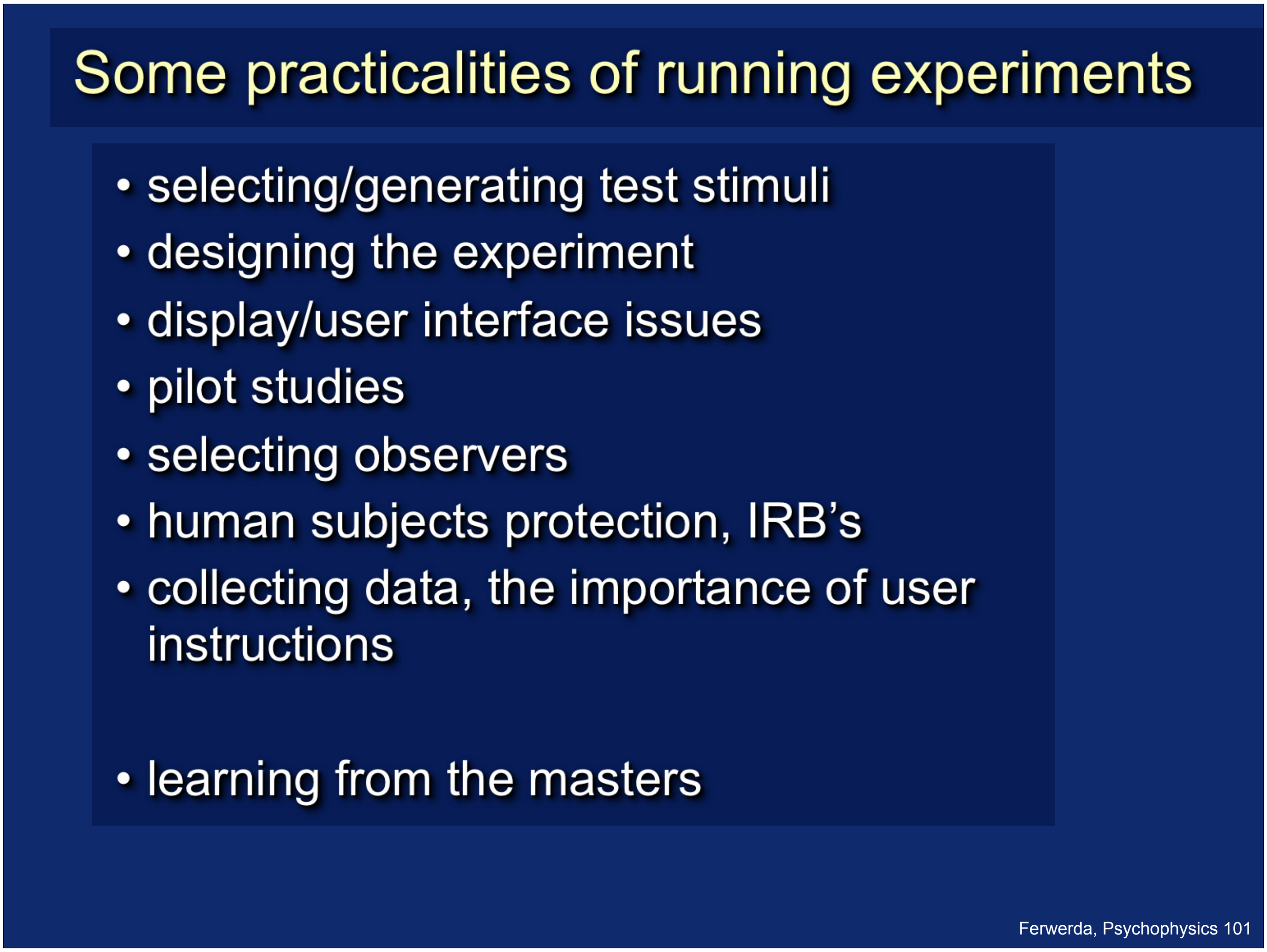
Summary

- scaling
  - four types of scales: nominal, ordinal, interval, ratio
  - indirect / direct scaling methods

		scale type			
		nominal	ordinal	interval	ratio
indirect	rating	X	X		
	pair comparison		X	X	
	ranking		X	X	
	category scaling		X	X	
direct	equisection				X
	magnitude est./prod.				X

- multidimensional scaling

- second topic was scaling
- four types of scales
  - nominal, ordinal, interval, ratio
  - different mathematical properties
- two main classes of scaling methods: indirect, direct
- table shows different scaling methods discussed, with the kinds of scales they're most often used to derive
- finally, multidimensional scaling: method for finding the psychophysical dimensions in complex stimuli



Some practicalities of running experiments

- selecting/generating test stimuli
- designing the experiment
- display/user interface issues
- pilot studies
- selecting observers
- human subjects protection, IRB’s
- collecting data, the importance of user instructions

- learning from the masters

- for threshold studies need to select stimuli that span threshold range
- for scaling studies need to select/generate stimuli that span a significant range of the parameter of interest
  - can generally be selected by eyeballing or informal pretesting
- when generating stimuli need to be sure that all variables/parameters but the ones being tested are held constant or controlled for
- need to select an experimental procedure that is appropriate to the stimulus set
  - discrete, continuous, uni/multi dimensional, size of stimulus set
- display viewing conditions must not interfere with/confound properties being tested/measured
- user input/response methods must not limit/impede user response
- running experiments is a time/labor intensive process for all involved
- much time/effort/mistakes can be saved by running small pilot studies to test/refine stimuli/design/procedure
- subject pool must be representative of find population of interest
  - undergraduate psychology majors!, vision researchers!
- many organizations require procedures to physically/legally protect human subjects
- human subjects committees, independent review boards (IRB’s)
- ignore at your own peril
- think long and hard about what task you’re going to ask your subjects to do and how you’re going to instruct them to do it
- data collected often depends critically on user instructions
- understand the effects of user learning, problem solving strategy development
- learn from the masters by closely reading methods/procedures sections of papers in high quality journals (J. of Vis., Vis. Res., J. Opt. Soc. Am., Perception, SIGGRAPH, ACM TAP,...)

Resources -books

- Gescheider, G.A. (1997) *Psychophysics: The Fundamentals*, 3rd Edition. Erlbaum.
- Bartelson, C.J. and Grum, F. (Eds.) (1984) *Optical Radiation Measurements, Vol 5: Visual Measurements.* Academic Press, New York.
- Engeldrum, P.G. (2000) *Psychometric scaling: A Toolkit for Imaging Systems Development*. Imcotek Press.
- Guilford, J.P. (1954) *Psychometric methods*. Mcgraw-Hill.
- Torgerson, W.S. (1960) *Theory and Methods of Scaling*. Wiley.
- Green, D.M. and Swets, J.A. (1966) *Signal Detection Theory and Psychophysics*. Wiley.
- Fechner, G.T. (1966) *Elements of Psychophysics.* Holt, Rinehart & Winston

Resources - papers/standards

- ASTM (American Society for Testing and Materials), Standard Guide for Conducting Visual Experiments, E1808-96
- ASTM (American Society for Testing and Materials), Standard Guide for Selection, Evaluation, and Training of Observers, E1499-94
- CIE Technical Committee 1-34 Testing Color-Appearance Models: Guidelines for Coordinated Research - Alessi, P.J. (1994) Color Research and Applications,19, 48-58.
- Use of computers and cathode-ray-tube displays in visual psychophysics - special issues of the journal Spatial Vision 10 (4) and 11(1) - http://www.hans.strasburger.de/

Resources - software

- Psychophysics Toolbox
  - Matlab-based - http://psychtoolbox.org/
- Psychophysica/Cinematica
  - Mathematica/Quicktime-based - http://vision.arc.nasa.gov/mathematica/psychophysica/
- Strasburger's overview of psychophysics software -
  - http://www.hans.strasburger.de/

Updates/Errata/Contact

- http://www.cis.rit.edu/jaf