Measurement Basics

Geoff Norman

Department of Clinical Epidemiology & Biostatistics
Program for Educational Research and Development

Agenda

- Scale construction
  - Issues in rating scale construction
  - Summing items to scales
- Measurement criteria
  - Reliability
  - Validity
  - Feasibility , acceptability
- Setting a pass score
- G Theory

When to use Rating scales

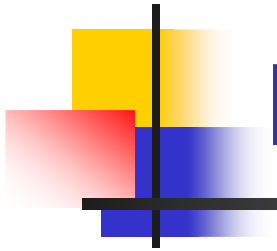
- Assessment of student performance
- Teaching ratings
- Course ratings
- Patient satisfaction
- Tutorial / peer / self ratings

Etc.

When do you require a scale?

Anytime there's no easily applied "objective" measurement tool

- Determining a patient's status
- Assessing Quality of Life
- Evaluating programs
- Performing cost/benefit analysis
- Documenting patient needs
- Assessing competence levels



Basic Characteristics of an Item

- > 7 categories or boxes (Miller 7+/-2)
- 4 or more descriptors
- Even / odd -- it depends

Example

Overall, how would you rate this workshop?

1	2	3	4	5	6	7
Worse than a fork in the eye		Delightfully unpleasant		Mildly entertaining		I heard angels speak

Would you recommend this section to a friend?

1	2	3	4	5	6	7
Only one my sister dated		Maybe an acquaintance		Without hesitation		I already registered him

Rating vs Checkist

- It appears that the more we can reduce competence to discrete, observable, countable behaviours, the more reliable and valid they will be.
- The appearance is WRONG!!
- Ratings are as reliable, more valid than checklists

Station 3: Renovascular hypertension

Student Name:_____

Instructions:
Please score the student's overall performance on this station by ticking one of the boxes below. In assessing performance the following are considered to be some of the important issues.

	Done Well	**Done Poorly**	**Not Done**
1. Checks blood pressure in left arm to see if readings are equal in both arms			
2. Visualizes optic fundi and looks for hemorrhages or exudates.			
3. Auscultation of the chest listening for crackles.			
4. JVP-looking for neck vein elevation.			
5. Character & location of apex.			
6. Cardiac auscultation to determine presence of S3 or S4.			
7. Palpation/auscultation over femoral pulses			
8. Auscultation of abdomen for bruits			
9. Palpation of abdomen for enlarged kidney			
10. Examination of periphery for edema			

Overall performance:

/20

Station 3: Renovascular hypertension

Student Name:_____

Instructions:
Please score the student's overall performance on this station by ticking one of the boxes below. In assessing performance the following are considered to be some of the important issues.

		Notes
	Checks blood pressure in left arm to see if readings are equal in both arms	
2.	Visualizes optic fundi and looks for hemorrhages or exudates.	
3.	Auscultation of the chest listening for crackles.	
4.	JVP-looking for neck vein elevation.	
5.	Character & location of cardiac apex.	
6.	Cardiac auscultation to determine the presence of S3 or S4.	
7.	Palpation/auscultation over femoral pulses	
8.	Auscultation of abdomen for renal bruits	
9.	Palpation of abdomen for enlarged kidney	
10.	Examination of periphery for edema Assess physical examination technique - systematic/non systematic, respect for patient, put patient at ease etc	

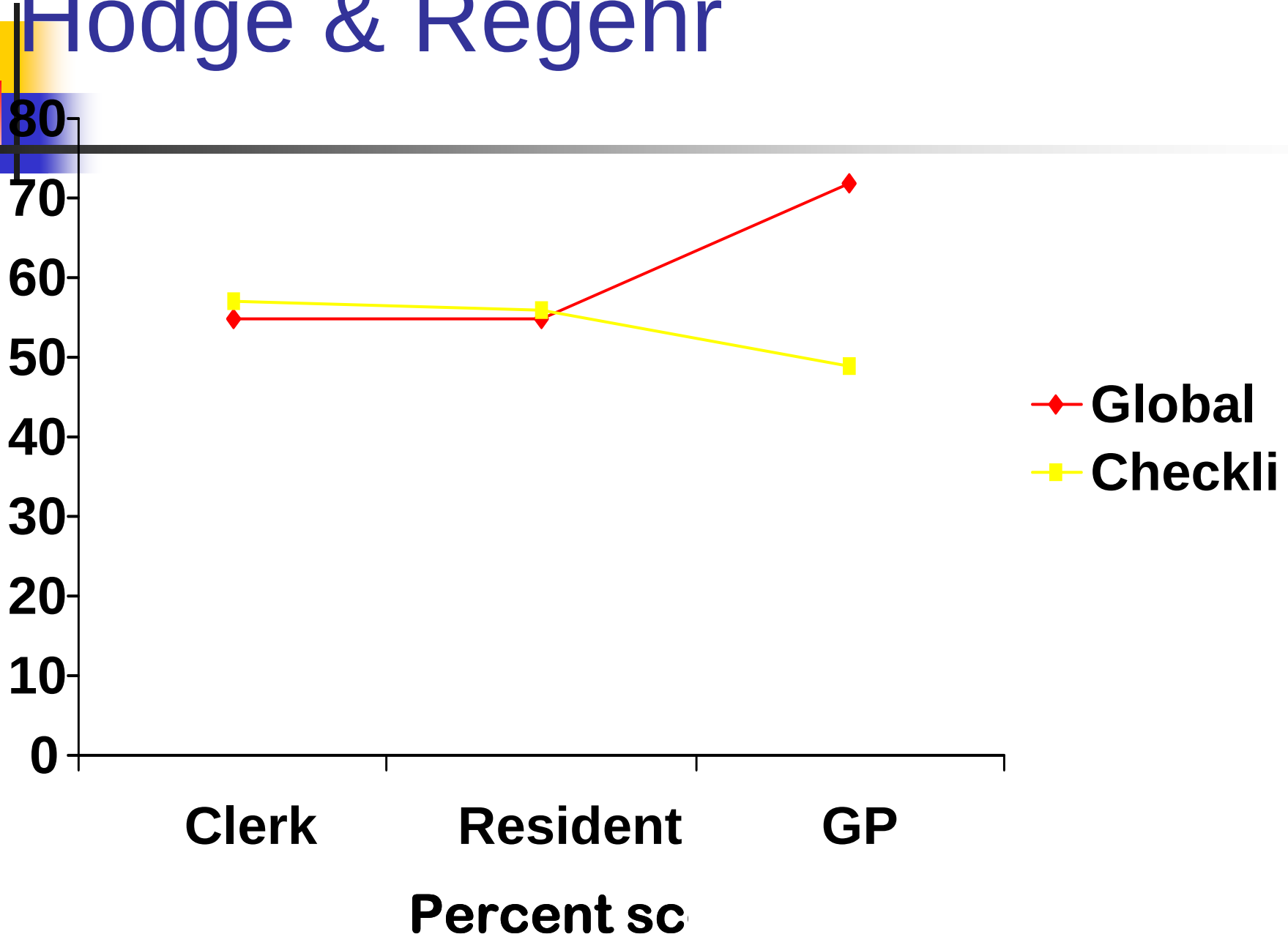
Overall performance:

[ ]	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
Unsatisfactory		**Borderline**		**Good**		**Excellent**

- Reliability
  - Inter-rater --- 0.7—0.8 (global or checklist)
  - Overall test (20 stn) – 0.8 (global > check)

- Validity
  - Against level of education
  - Against other performance measures

Hodge & Regehr



Item --> Scale

- Many complex strategies to weight items in arriving at a score:
  - Expert judgement
    (4 x Item a + 2 x item B)
  - Regression weights
    (predicting some criterion)
  - Unintentional weights
    (Physical fn -- 12 items, Social fn. 3 items)
  - Multiplicative Weights (Individualized)
    (Importance x Ability)

- Complex weighting schema almost never have any superiority to an equal weighting approach

"On Estimating Coefficients in Linear Models:
It Don't Make No Nevermind" (H Wainer, 1976)

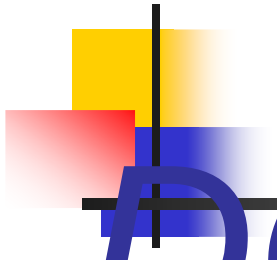
"The Robust Beauty of Improper Linear Models"
(R Dawes, 1968)

Scale >> Category

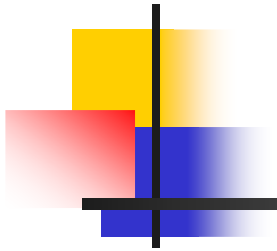
(the ubiquitous 2 x 2 table)

- When you're looking at 2 groups, it becomes irresistible to develop a cutoff like:
  - Hypertension = SBP>160 and DBP>95
  - Hypercholesterolemia = LDL>130
  - Depression = CES-D > 22
  - Improved / Stable / Worse Quality of Life
    (e.g. > 1 MID)

and classify people as diseased or not with / without Treatment

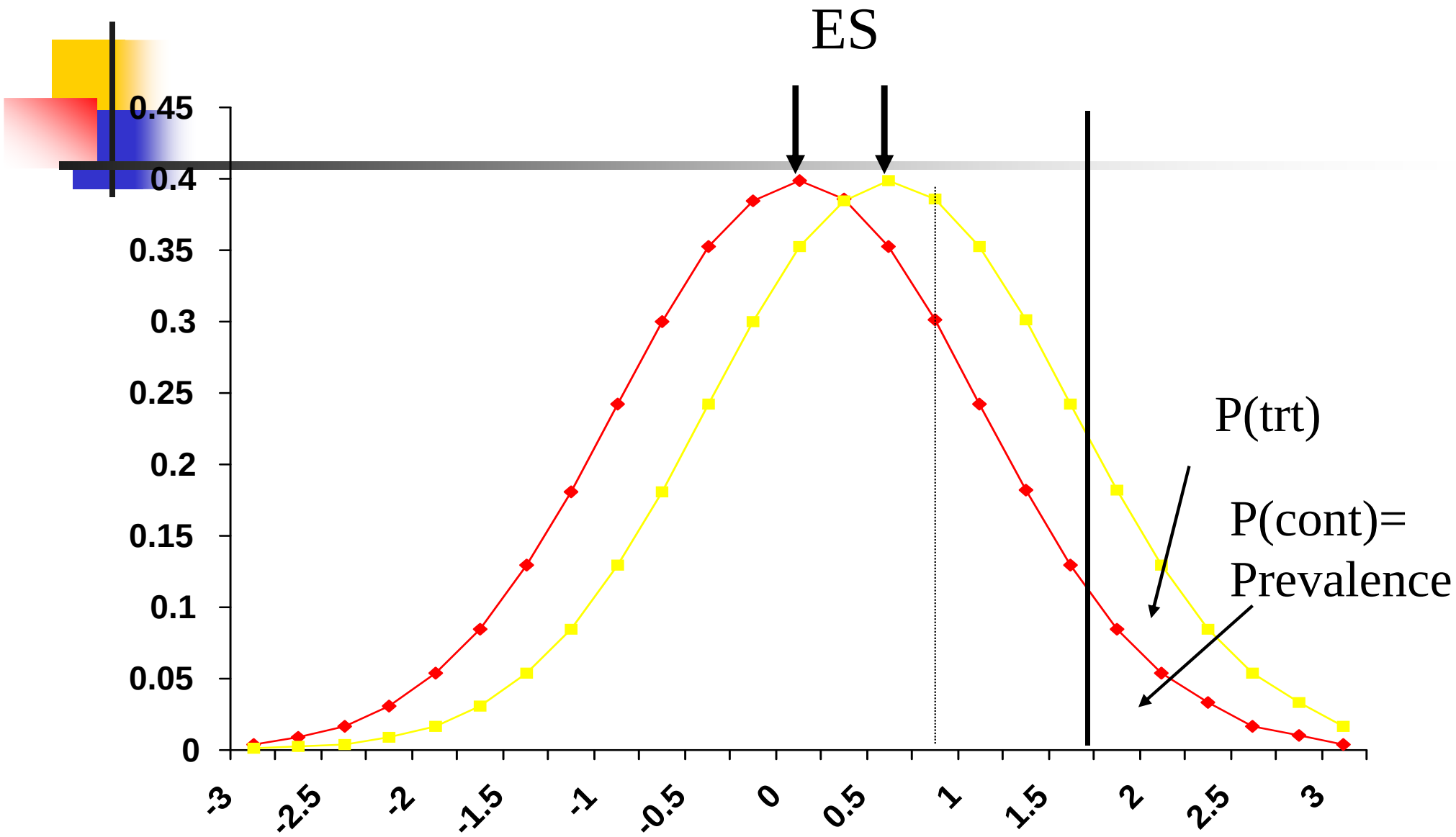


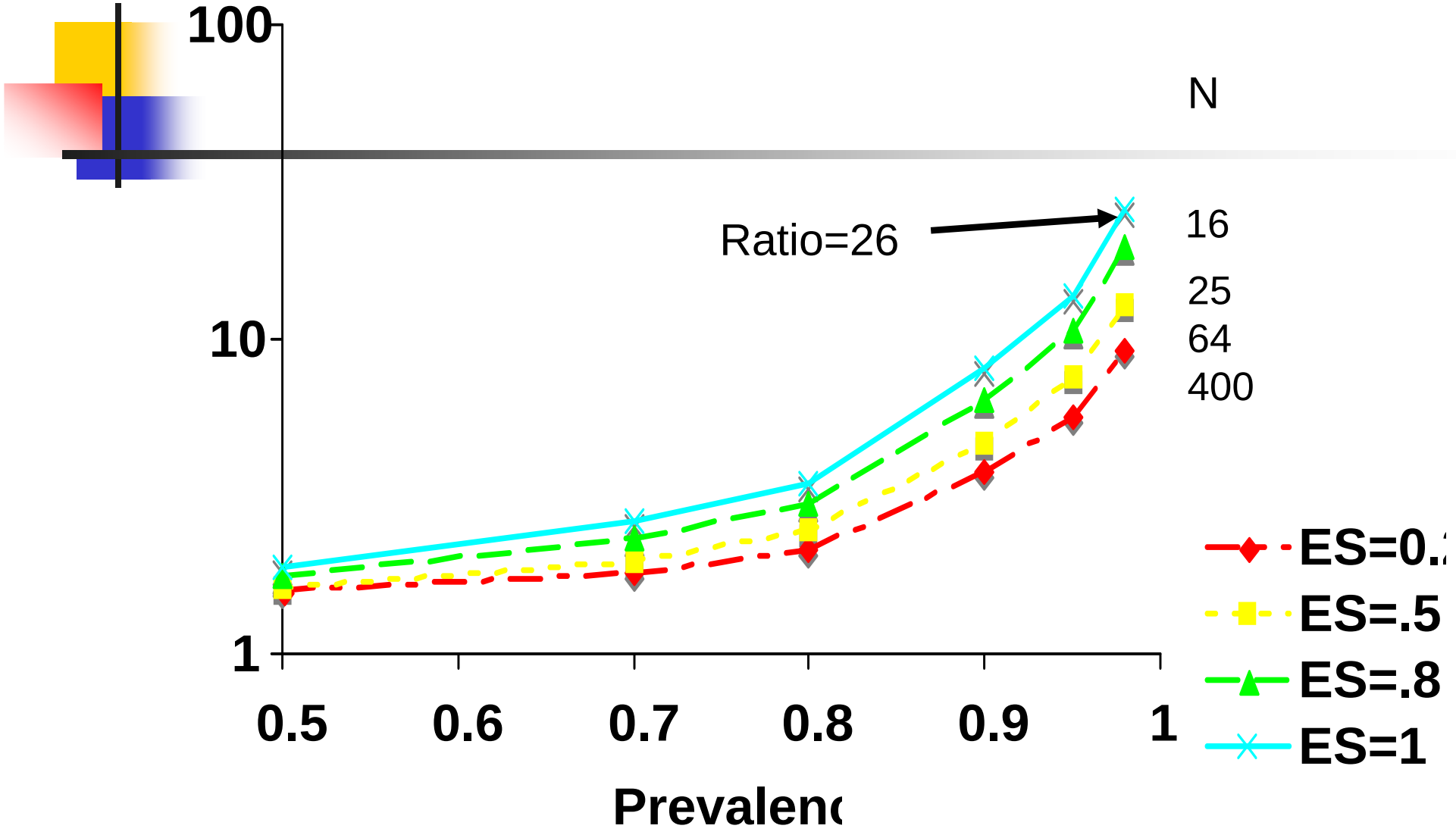
DON'T DO ≠ IT!!



- dichotomizing continuous variables may cost statistical power and sample size

  from 1.5 to 25 depending on cut point


ES
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
P(trt)
P(cont)=
Prevalence


100
10
1
N
Ratio=26
16
25
64
400
0.5
0.6
0.7
0.8
0.9
1
Prevalenc
ES=0.
ES=.5
ES=.8
ES=1

The Four “-itys” of Good Measurement

(unabashed plagiarized from Kevin Eva, with permission)

What makes a good scale?

Reliability

Validity

Feasibility

Acceptability

Reliability

"Measurements of individuals on different occasions, or by different observers, or by similar or parallel tests, produce the same or similar results."

Streiner & Norman, Health Measurement Scales (2nd ed.)

Why Reliability is Important:

- Using an unreliable tool:
  - Is equivalent to measuring your height with a spring
  - Creates a source of error variance
    - May not see real difference (Type II error)
    - Need more participants to show a difference
  - Places an upper limit on validity
    - If it’s not reliable, it’s measuring noise so it can’t be valid

Classic Reliability Theory

Raw Score = True Score + Error

Sources of Error:

Misinterpretation

Biases (e.g., extreme boxes)

Inexperience

Inter-rater differences

Et cetera, et cetera …

Reliability

Tests constructed to differentiate between subjects (people or something else)

$$\text{Reliability} = \sigma^2_{True} / (\sigma^2_{True} + \sigma^2_{Error}/n)$$

Implications:

Reliability not a fixed property of scale

More heterogeneity = more reliability

More observations =more reliability

A test that doesn't discriminate is useless

What's wrong with just error?

"The error of measurement in this thermometer is +/- 2 deg. Celsius"

Is this a good thermometer or a bad one?

Ans: It depends……

… on the true variation in the thing you're trying to measure

Classical Test Theory (ca. 1910)

Observed = True Score + Error

Reliability = “the proportion of the variance in the scores contributed by true differences among people”

$$\text{Reliability} = \frac{\text{True variance}}{\text{(True var + Error var)}}.$$

$$= \frac{\text{(Total variance - Error variance)}}{\text{Total variance}}$$

A Typical Design

Subject	Rater 1	Rater 2	Rater 3	Average
A	5	4	6	5.0
B	4	6	8	6.0
C	2	4	6	4.0
D	3	5	4	4.0
E	3	3	3	3.0
F	1	2	3	2.0
Avge	3.0	4.0	5.0	4.00

A Typical Design

Subject	Rater 1	Rater 2	Rater 3	Average
A	5	4	6	5.0
B	4	6	8	6.0
C	2	4	6	4.0
D	3	5	4	4.0
E	3	3	3	3.0
F	1	2	3	2.0
Avge	3.0	4.0	5.0	4.00

A Typical Design

Subject	Rater 1	Rater 2	Rater 3	Average
A	5	4	6	5.0
B	4	6	8	6.0
C	2	4	6	4.0
D	3	5	4	4.0
E	3	3	3	3.0
F	1	2	3	2.0
Avge	3.0	4.0	5.0	4.00

Subjects

Raters

A Typical Design

Subject	Rater 1	Rater 2	Rater 3	Average
A	5	4	6	5.0
B	4	6	8	6.0
C	2	4	6	4.0
D	3	5	4	4.0
E	3	3	3	3.0
F	1	2	3	2.0
Avge	3.0	4.0	5.0	4.00

Error

Subjects

Raters

Repeated Measures ANOVA

Source	Sum Squares	d.f.	Mean Sq	F	p
Subjects	16.94	5	3.39	--	--
Raters	4.78	2	2.39	1.79	.23
Error	13.89	10	1.39		

Variance Components

- Subjects 0.67
- Raters 0.17
- Error 1.39

$$\text{Reliability} = \frac{0.67}{0.67 + 1.39\ (+0.17)} = 0.30$$

Types of Reliability

1. Internal Consistency
   - Are the items measuring the same thing?
2. Test - Retest
   - Do you get the same results on two occasions?
3. Intra - rater
   - Does one rater give the same results twice?
4. Inter - rater
   - Do two raters agree with each other?

Measuring Reliability

1. Internal Consistency
   - Split-halves - correlation between randomly divided halves
   - Kuder-Richardson 20 - useful if dichotomous
   - Cronbach's α - average of all split half reliabilities

2. Test - Retest, Intra - Rater, and Inter - rater
   - Pearson's r
   - Intra-Class Correlation (ICC)
   - Kappa, Weighted Kappa

Cronbach's Conundrum

(Two Disciplines of Scientific Psychology Revisited)

- When you're looking at reliability, more variation = more reliability

  (Correlational)

- When you're doing an experiment, more variation = less significant treatment effect

  (Experimental)

Validity

- Is the test measuring what we think it's measuring?
- Not guaranteed by reliability
- Determines legitimacy of conclusions drawn based on scores derived from scale

Why Validity is Important:

- Using an invalid tool:
  - Is equivalent to measuring your height with a blood pressure cuff
  - Leads to wrong conclusions
    - False positive
    - False negative
  - Is plainly unethical in many situations

Types of Validity

1. Face Validity
   - Does it appear to measure what you think it does?
2. Content Validity
   - Does it tap all relevant areas and no irrelevant areas?
3. Criterion Validity
   - Are the results consistent with other measures?
4. Construct Validity
   - Can we predict differences based on constructs?

Types of Validity

1. Face Validity

- Does it appear to measure what you think it does?
- The items presented should appear to be similar to those whose ‘real life’ performance is being predicted
- “On the face of it” does it appear valid

Types of Validity

2. Content Validity

- Does it tap all relevant areas and no irrelevant areas?
- Define important areas based on
  - Existing tools, Clinical observation, Expert opinion, Research, Theory, Patients' reports
- Check with experts from wide variety of areas (not just colleagues)

Types of Validity

3. Criterion Validity

- Are the results consistent with other measures?
- Concurrent - Examine relationship between criterion measure and scale at time of administration
- Predictive - Examine relationship between scale and future outcome

Types of Validity

4 Construct Validity

- Can we predict differences based on constructs?

- Test measure against theory
  - Change with therapy (Responsiveness)
  - Correlation with other measures
  - No correlation with other measures (Discriminant)

Measuring Validity

- There is no single measure of validity

- Validity testing typically requires systematic, often long-term, testing

- Proper test dependent on type of study
  - Extreme groups - t-test
  - Change - two way ANOVA
  - Criterion - Pearson's Correlation

Feasibility

- A measure is only useful to the extent that it can be used

- Avoid ambiguity
  - Not double-barreled
  - Don’t use negative wording
  - Avoid jargon

- Pay attention to reading level

Feasibility

- How much training is required?
- How easy is it to score?
  - Avoid weighting responses
  - Be aware of unintentional weighting
- What does it cost to administer?

Acceptability

- A measure is only useful to the extent that it will be used
- Be brief
  - Remember though that reliability increases as a function of number of items on the scale
- Use only items for which there is a variety of responses

Acceptability

- Be aware of social desirability bias

- Other biases:
  - Yay or nay saying
  - End aversion
  - Halo effect
  - “Lake Wobegone” effect

Determining the Cut Score

Four Broad Approaches

- Fixed
  - The pass mark is 60%
    (regardless of the difficulty of the exam)
- Norm- referenced (Relative)
  - The failure rate is 5%
    (regardless of the ability of the class)
- Criterion Referenced (Absolute)
  - The eventual failure rate is determined by the characteristics of the test
- Combination
  - Borderline group
  - Hofstee

FIXED APPROACHES

Historically dominant

Logically indefensible

- Test difficulty varies
- Candidate ability varies

(remember IRT methods)

Relative Methods
(Norm Referenced or fixed percentage)

- Decision to pass or fail based on relative rank in the class:
  - e.g. Bottom 5% of class will fail
- Actually may be defensible with **large** numbers of candidates
  - 2000 candidates may vary less year to year than 200 items

Absolute (criterion referenced)

Decision is based on a judgment item by item:

- Angoff
- Nedelsky

The capital of Canada is:

a) Toronto

b) Montreal

c) Washington

d) Ottawa

e) London

ANGOFF

- Think about a “borderline group” of European high school students (a group with a 50% chance of failing the geography exam overall)

- What proportion of them would get this item right?

Nedelsky

- How many of the wrong answers might a borderline student select?
- What is a “forgiveable” wrong answer

  (e.g. Ottawa is correct, Toronto and Montreal are acceptable)

No. acceptable wrong	Contribution to pass mark
0	1.0
1/5	.8
2/5	.6
3/5	.4
4/5	.2
5/5	0

Advantages

- Based on absolute standards of performance
- Independent of particular candidate ability

Disadvantage

- Large inter-judge variation
- Unless actual performance at item level is available, pass rates can be wildly off
- How do you know if you're right?

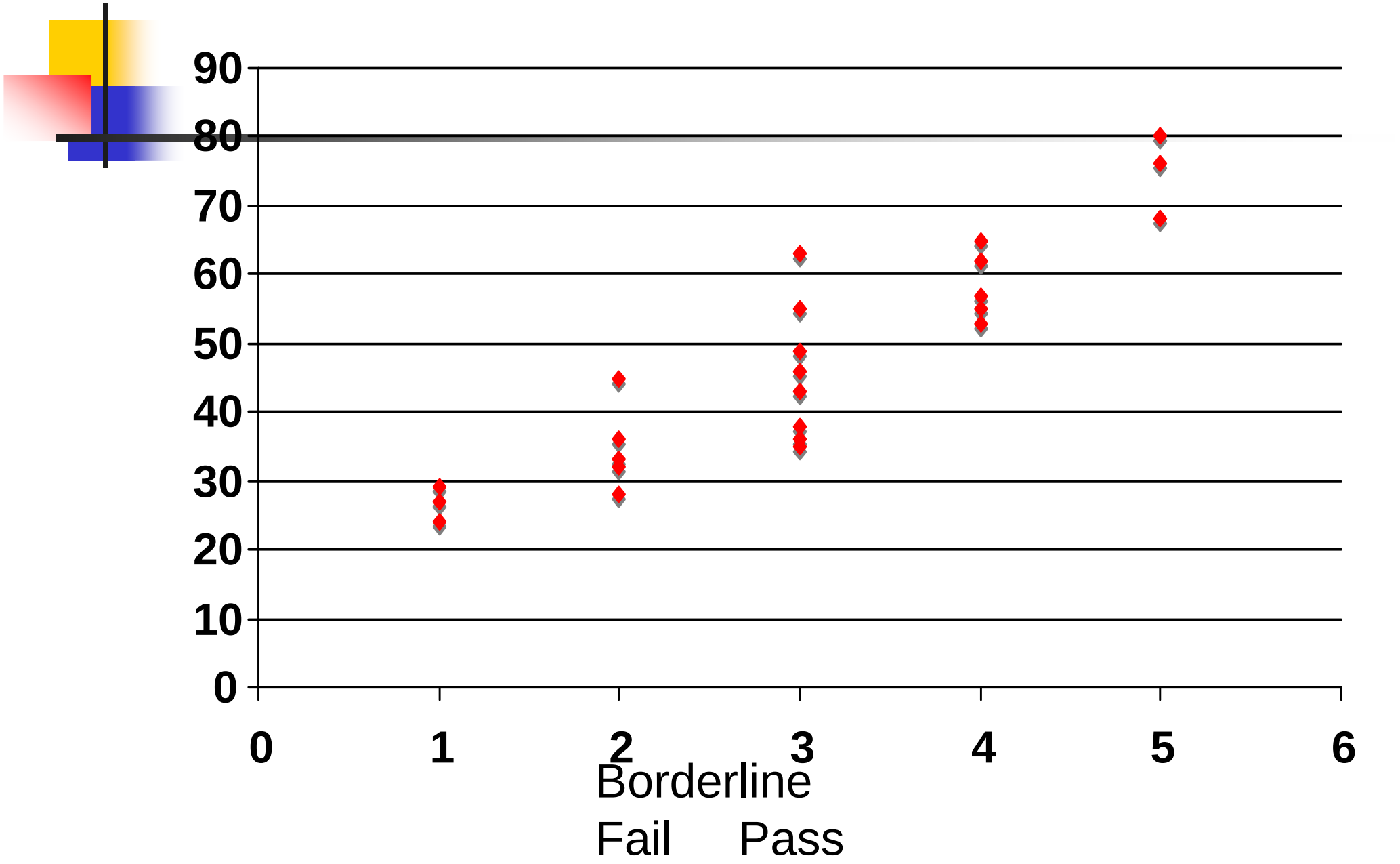
Borderline /Contrasting groups

Requires two kinds of data:

- Overall pass / fail decision / judgement
- Specific performance on test

e.g. in an OSCE, judge completes checklist and also rates overall performance as:

Clear Fail
Borderline Fail
Borderline Pass
Clear Pass
Excellent


90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
Borderline
Fail
Pass

Method 1 - Average of BGs

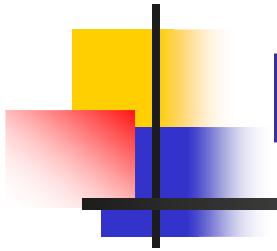
Bordeline fail:

32,36,45,33,28,35

Borderline Pass:

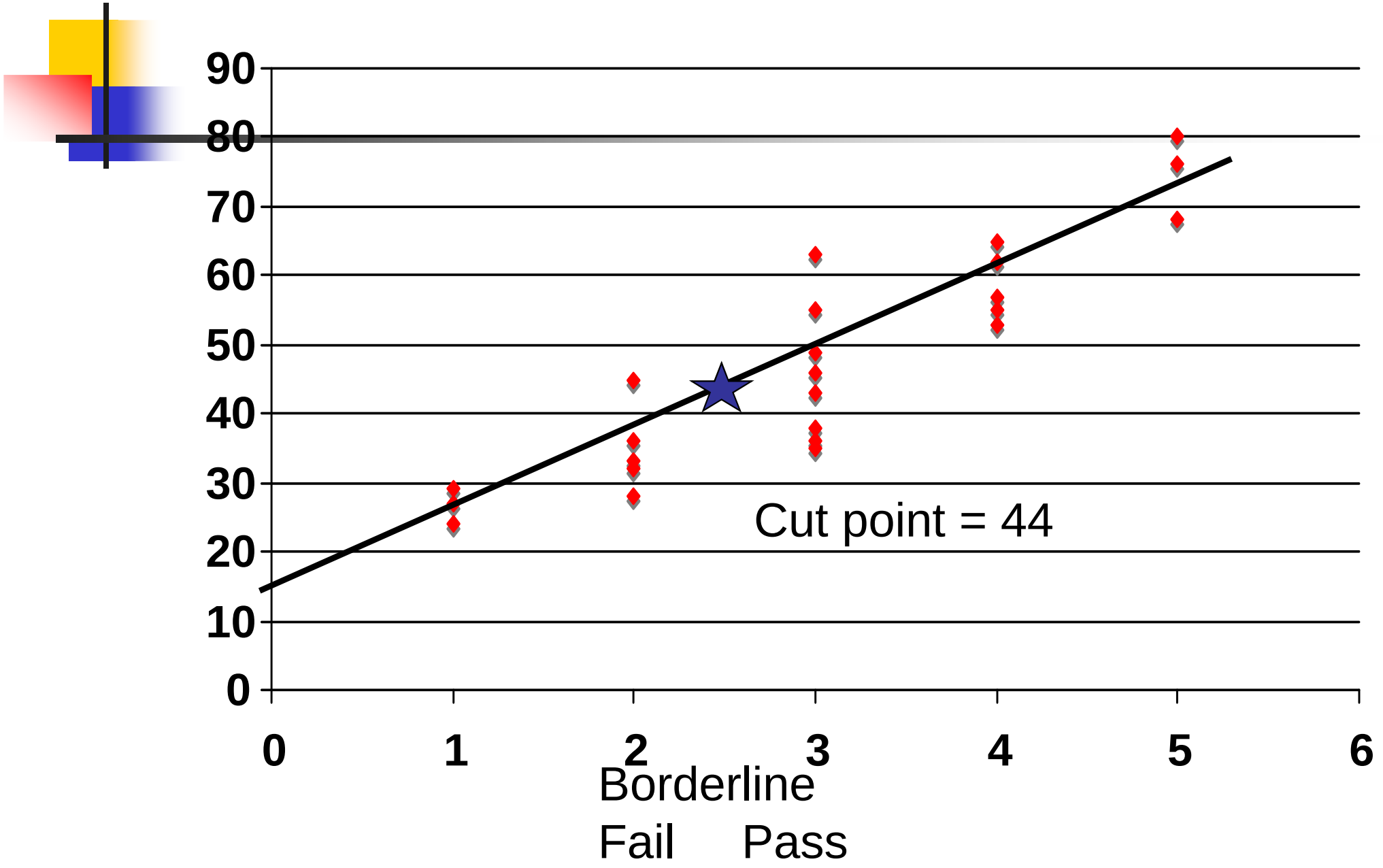
38,43,46,36,55,63,49,

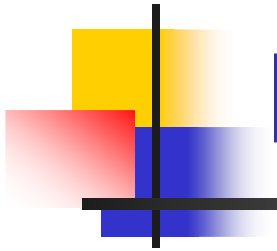
Mean = 41.46 = passmark



Method 2 - Regression

- Fit regression line to ALL the data
- Cmpute point on the line corresponding to X = 2.5


90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
Cut point = 44
Borderline
Fail
Pass



Method 1 vs Method 2

Method 2 uses all the data, so better for small n

Since there are always more BPs than BFs (from normal distribution) Method 1 is inherently biased

But the world uses it more anyway

Hofstee Method

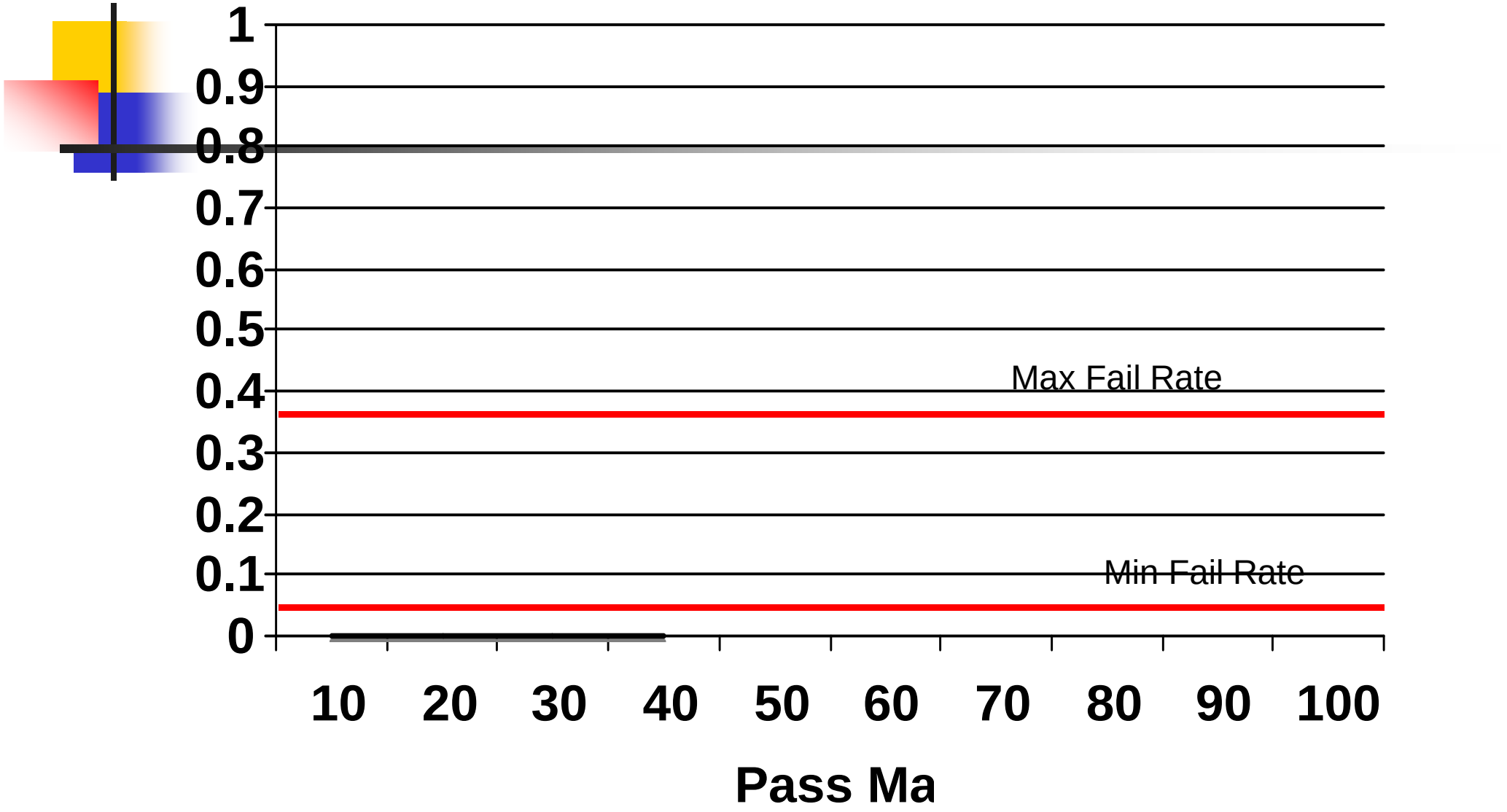
Judges must decide:

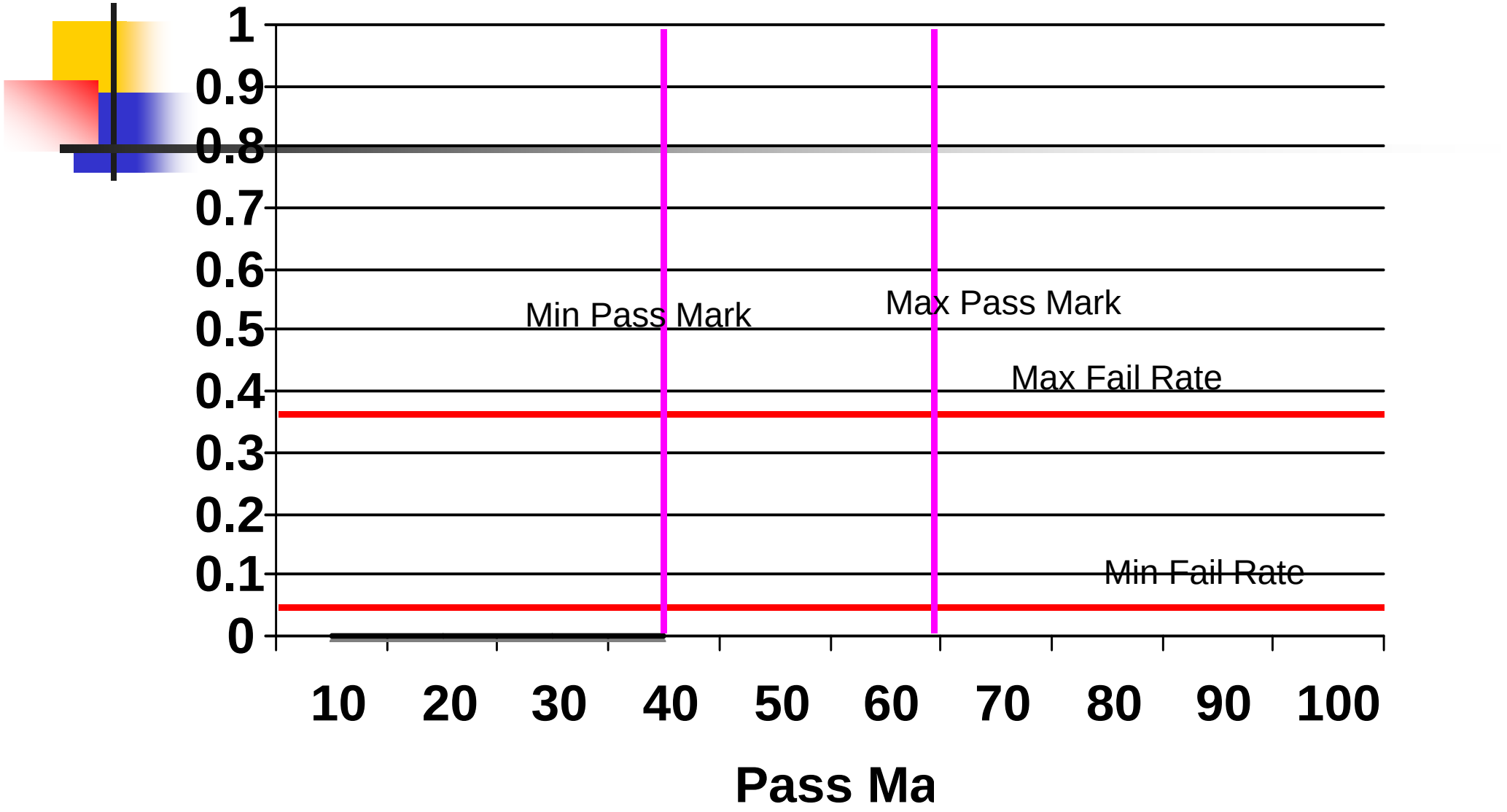
What is the *highest* acceptable pass mark? (could you fail someone with 75%)

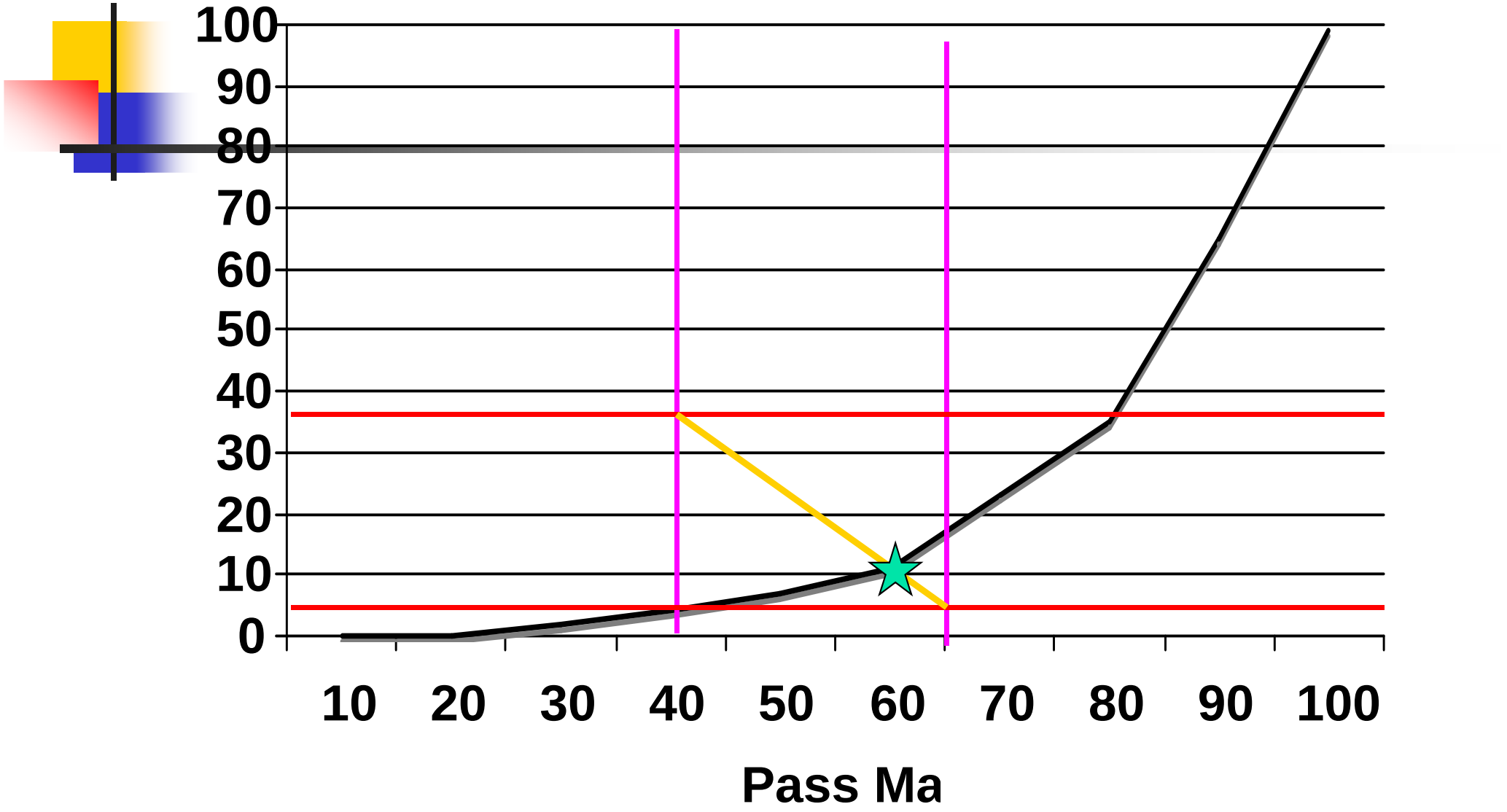
What is the *lowest* acceptable pass mark? (could you pass someone with 25%)

What is the *highest* acceptable failure rate? (could you fail 3/4 of the class)

What is the low*est* acceptable pass mark? (could you pass everyone?)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Max Fail Rate
Min Fail Rate
10
20
30
40
50
60
70
80
90
100
Pass Ma


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Min Pass Mark
Max Pass Mark
Max Fail Rate
Min Fail Rate
10
20
30
40
50
60
70
80
90
100
Pass Ma

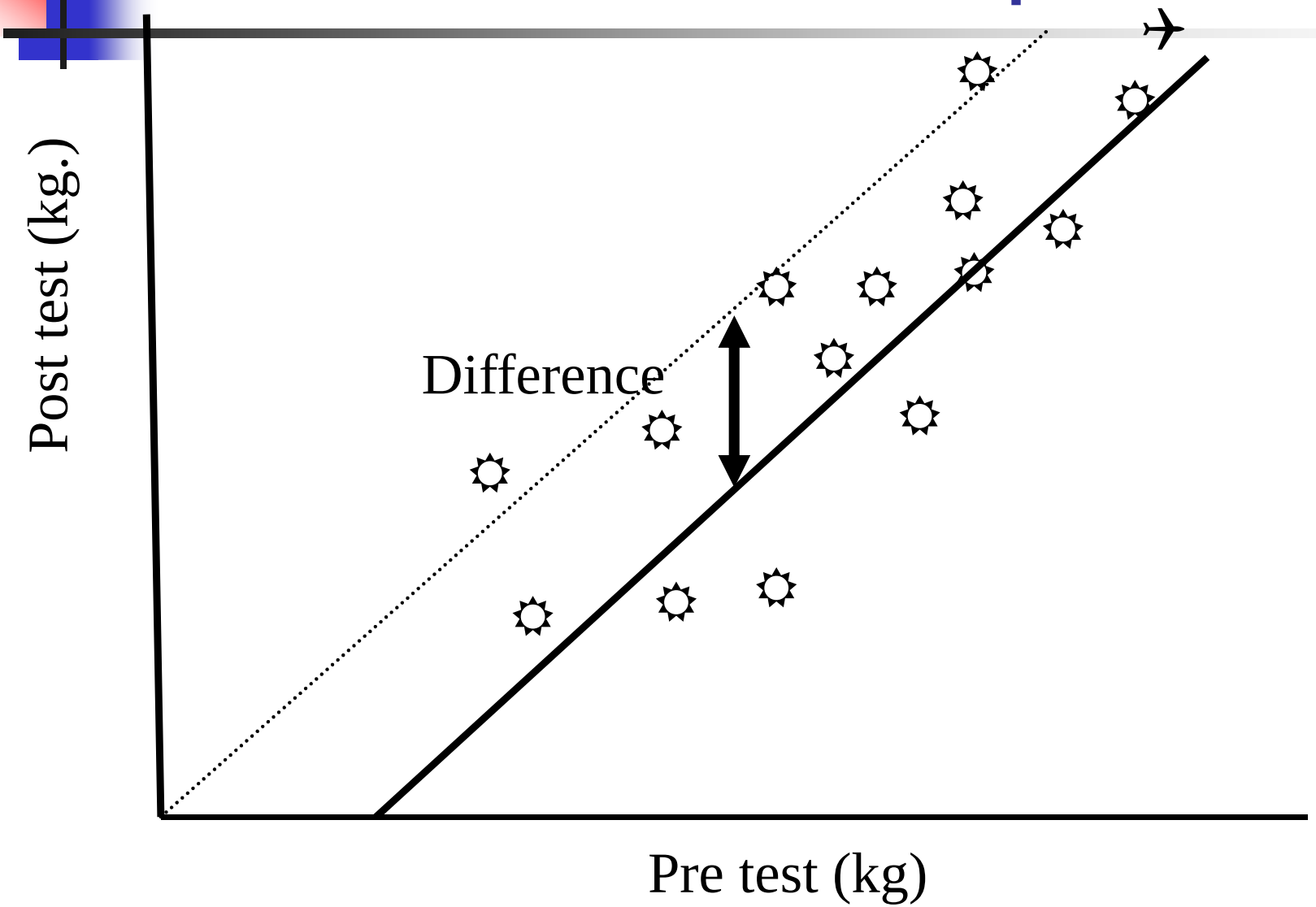

100
90
80
70
60
50
40
30
20
10
0
10
20
30
40
50
60
70
80
90
100
Pass Ma

SUMMARY

- Absolute fixed methods are indefensible
- Relative (norm ref) methods practical, and defensible on psychometric grounds
- Crit ref methods politically correct, but constely and may be inaccurate
- Compromise methods efficient, “authentic”

Are difference scores the best approach?

Pre and Post Diet plan



- But with measurement error, people are on the extremes (high , low) because:

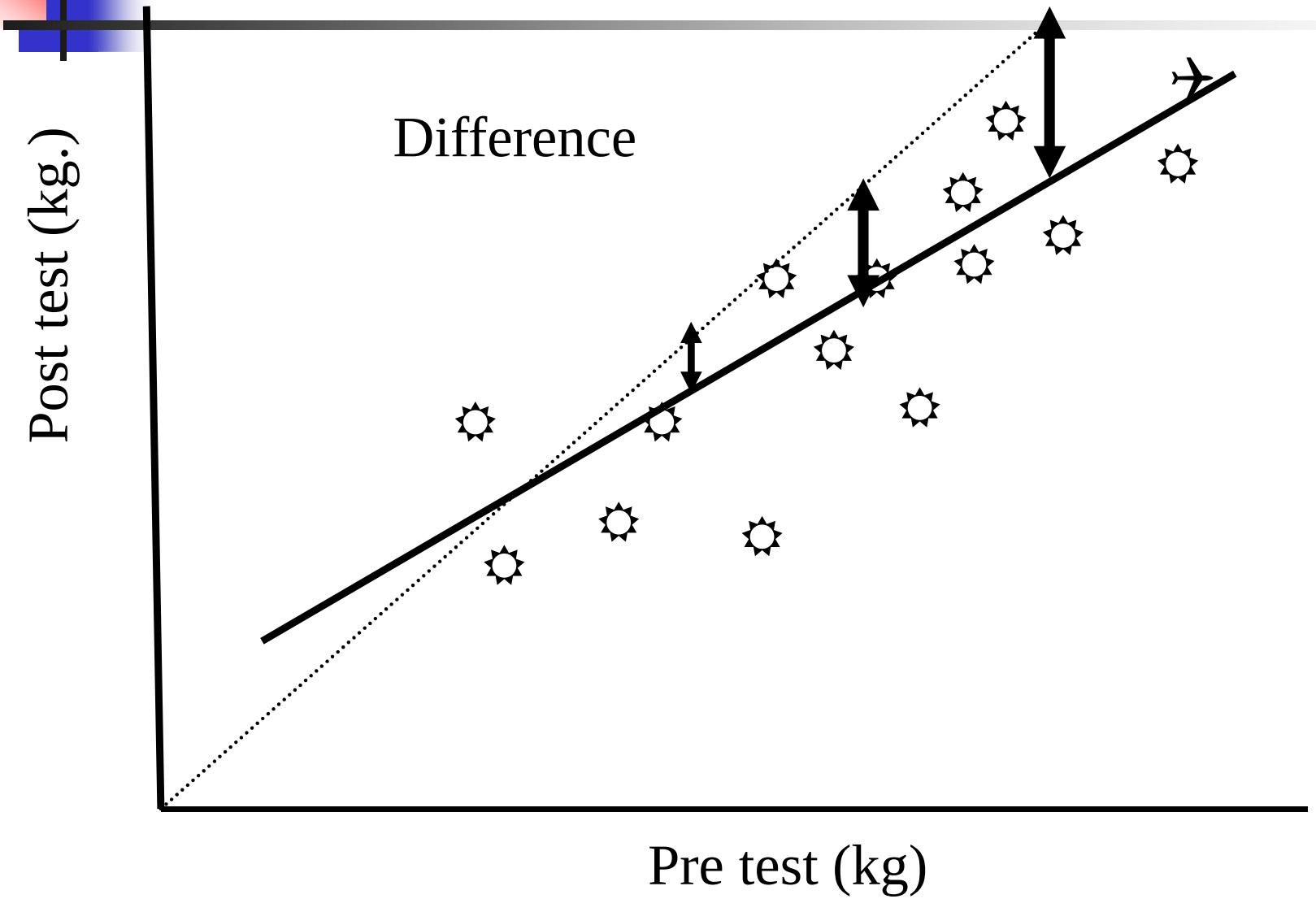
  a) They really are extreme

  b) Measurement errors conspired to put them there

  and when you test them again, they'll be less extreme

  (Regression to the Mean)

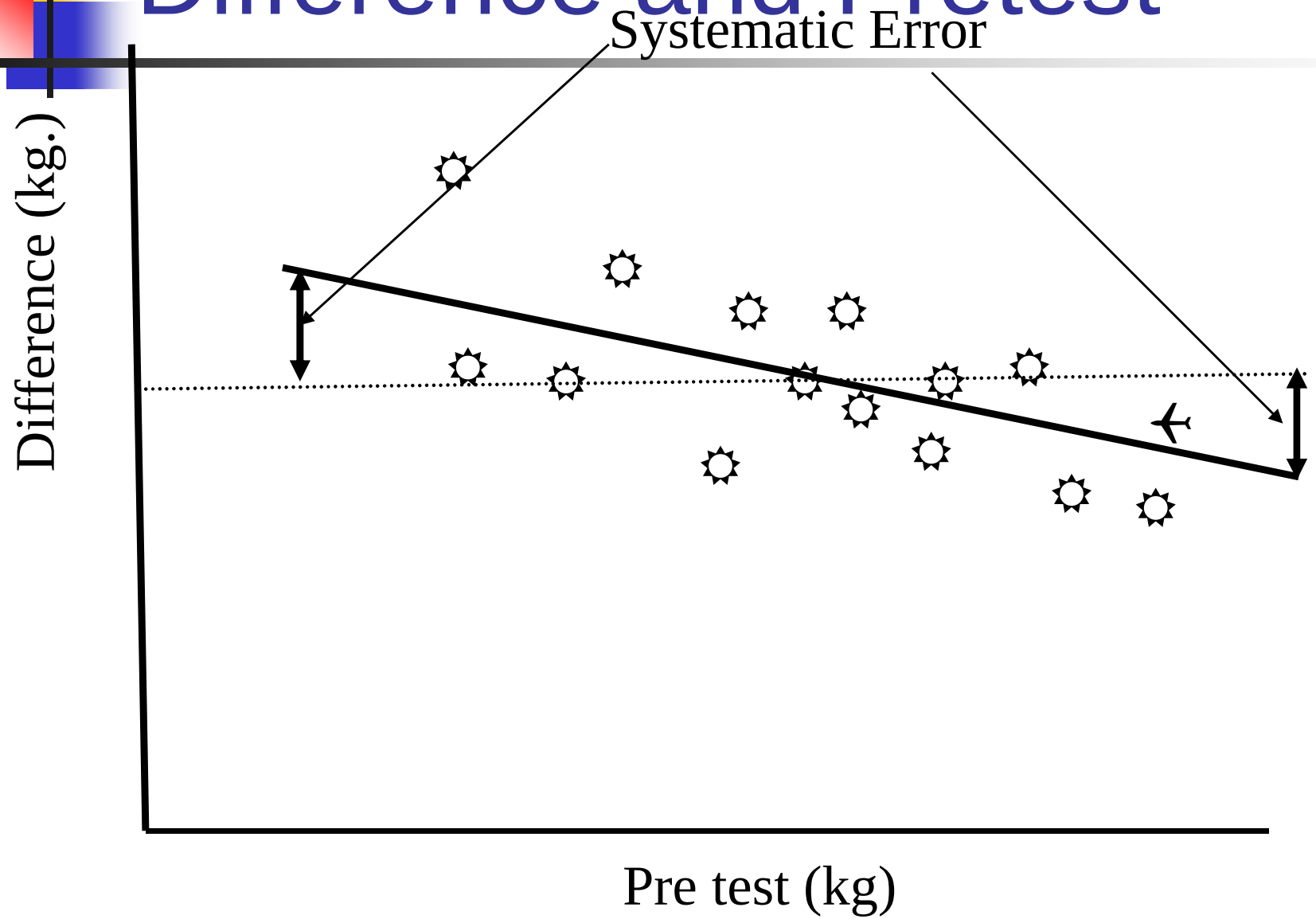
Pre and Post Diet plan



- Even if people all changed the same amount on average (except for random error)

  use of Difference Score introduces additional error

Difference and Pretest



The Solution

Difference Score:

Treatment = Posttest - Pretest

ANCOVA:

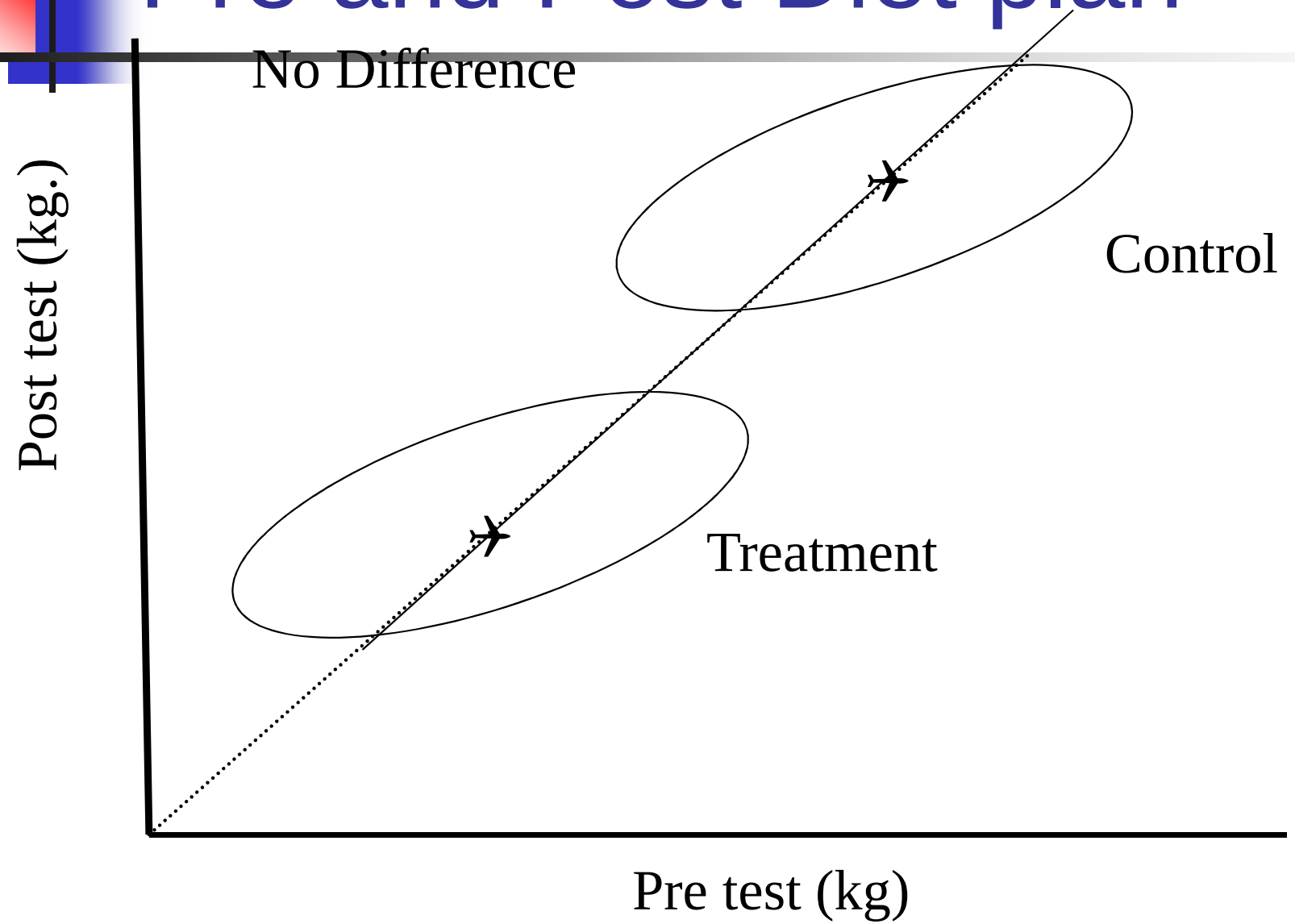
Treatment = Posttest - α (Pretest)

where α (< 1) estimated from the data

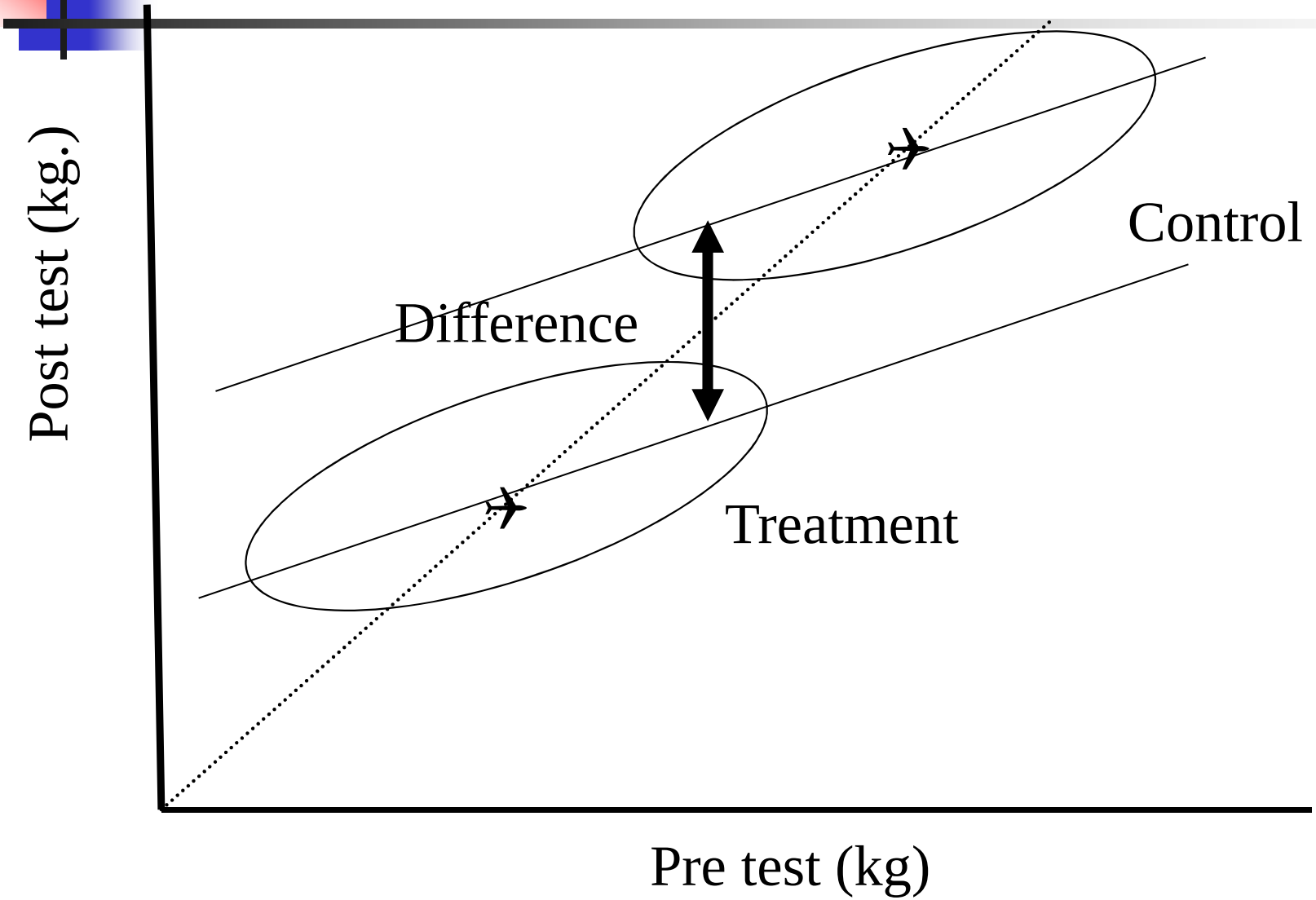
Should you always Pretest?

- In Program experiments, pretest can "give the game away" and becomes part of the treatment
- In cohort studies, pretest (or anything else) cannot be used to correct for baseline differences

Pre and Post Diet plan



Lord's Paradox



- In randomized experiments, pretest can be used to increase statistical power

…… but it may decrease power

- Post test only:

  Error (within) = σ^2 (patient) + σ^2 (error)

- Difference -- Posttest - Pretest

  Error (within) = σ^2 (error) + σ^2 (error)

If 2 σ^2 (error) > σ^2 (patient) + σ^2 (error)

If $2\,\sigma^2$ (error) > σ^2 (patient) + σ^2 (error)

…. σ^2 (error) > σ^2 (patient)

….. Reliability < 0.5

then Pretest reduces precision

Generalizability Theory:

A new approach to calculating reliability

(well, not really new. Cronbach, 1972)

So what's wrong with Classical Test Theory?

a) COMBINING DATA

- Imagine a 10 station OSCE
- With 2 raters (examiner, patient)
- And 4 scales:

  Knowledge, problem-solving, data gathering, interpersonal)

b) OPTIMIZING RELIABILITY

- Study 1: Inter-rater reliability = 0.65
- Study 2: Internal consistency =0.8
- Study 3: inter-case reliability = 0.4

Should we have:

2 raters, 4 scales, 10 stations?

4 raters, 5 scales, 4 stations?

1 rater, 2 scales, 40 stations?

- There are
  - 40 estimates of inter-rater reliability
  - 20 estimates of internal consistency
  - 8 estimates of inter-case reliability

And no defensible way to average them

- What we need is some way to analyze all the data in one study, compute true and error variances, and develop strategies to determine optimal ways to combine factors

- And that's what Generalizability Theory does…..

Instead of..

"what is the test-retest reliability….."

We say…

"to what extent can I generalize an observation of X to a different (rater, time, item)"

- The Facet
  - One of the variables or factors in the study design
- The “Universe score”
  - The average score over all possible levels of all facets (The “universe of observations”)

- The “facet of differentiation”
  - the object you are measuring (e.g student)
- The “facet of generalization”
  - the factor you are generalizing over (e.g.rater)
- The “fixed facet”
  - The factor you are holding constant
  - (e.g. time)

The Basic Approach

- Determine the major potential sources of variance
  - Rater, time, case, item…..
- Design a study which includes all these sources of variance in a factorial ANOVA design
- Compute SS, MS, variance

The MMI Study (admissions OSCE)

Sub	Case A		Case B		Case C			Case F	
	R1	R2	R3	R4	R5	R6	….	R11	R12
1	i1 i2 i3	i1 i2 i3							
2									
.									
.									
.									
18									

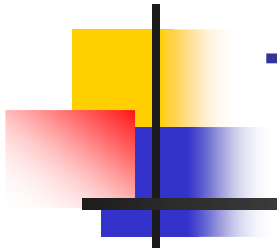
The Facets

- Subjects (S)
- Raters (R)
- Items (I)
- Cases (C)

Classical Equivalent	S	R	I	C
Inter-rater reliability	D	G	F	F
Internal Consistency	D	F	G	F
Inter-case	D	F	F	G
Overall test	D	G	G	G

Course in Critical Thinking

S	TIME 1					TIME 2			
a	i1	i2	…	i25		i1	i2	…	i25
b									
c									
d									
e									
f									
g									
h									
i									



The Facets

- Subjects (S)
- Items (I)
- Time (T)

Classical Equivalent	S	I	T
Internal Consistency	D	G	F
Test-Retest	D	F	G
Overall test	D	G	G
Change T1--> T2	G	F	D

A Simpler Study

- A course in writing. Essay at beginning and end. Three raters. 10 students.

		Time 1			Time 2			
STUD	R1	R2	R3	R1	R2	R3		
1	10	9	7	8	8	9	8.50	
2	7	5	8	6	8	7	6.83	3
	4	5	6	5	6	7	5.50	
4	8	7	5	6	7	6	6.50	
5	2	3	5	4	3	2	3.17	.
.								
.								
.								
10	4	3	2	3	4	3	3.17	

Source	SS	df	MS
Student	205.93	9	22.88
Time	0.0	1	0.0
S x T	3.00	9	0.333
Rater	0.033	2	.0167
S x T	22.97	18	1.276
T x R	4.30	2	2.15
Error	20.70	18	1.15

Rule 1: Every MS contains the σ^2 for that effect + σ^2 for all higher order interactions with that effect

Rule 2: Each σ^2 is multiplied by the number of levels of all factors not contained in theσ^2.

$$MS(err) = V(err) = 1.15$$

$$MS(SxR) = V(err) + t\,V(SxR) = 1.276$$

$$MS(SxT) = V(err) + r\,V(SxT) = 0.333$$

$$MS(S) = V(err) + t\,V(SxR) + rV(SxT) + rtV(S)$$

Rule 3: To compute V from MS, begin with MS for the effect, then subtract first order interactions, add second order interactions, subtract 3rd order…
then divide by the multiplier of the V (Π of all n's excluded)

MS(err) = V(err) = 1.15

MS(SxR) = V(err) + t V(SxR) = 1.276

V(SxR) = (MS(SxR)-MS(err)/2 = (1.276-1.15)2= .061

MS(SxT) = V(err) + r V(SxT) = 0.333

V(SxT) = (0.333-1.15)/2=0

MS(S) = V(err) + t V(SxR) + rV(SxT) + rtV(S)

V(S) = (22.88-.333-1.276+1.15)/(3x2) = 3.73

Rule 4: If V < 0, set V = 0

V(SXR) = .061

V(SXT)=0

V(S)= 3.73

V(RXT)=0.10

V(R) = 0.0

V(T)=0.0

Computing G coefficients

G = {V(total) - V(error)} / V(total)

Rule 5:

Denominator contains effect of interest (facet of diff) and all higher order interactions with the facet

Rule 6:

Numerator excludes interaction with facet of generalization, so contains facet of diff and all interactions with FIXED facets

To what extent can I generalize from an observation made by any rater, to one made by any other rater (inter-rater R)

- Diff = Student, Gen = Rater, Fix = Time

$$G = \frac{S + SxT}{S + SxR + SxT\ (+/o\ T, R) + err}$$

$$= \frac{3.73 + 0}{3.73 + 0 + .061 + 1.15} = 0.75$$

To what extent can I generalize from an observation made by any rater, to one made by the same rater on another time (intra-rater R)

- Diff = Student, Gen = Time, Fix = Rater

$$G = \frac{S + SxR}{S + SxR + SxT\ (+/o\ T, R) + err}$$

$$= \frac{3.73 + 0.061}{3.73 + 0 + .061 + 1.15} = 0.77$$

Rule 8: When computing G of *average* score across n levels, error variance is divided by “n”

How reliable is the score based on 3 raters and 2 times (inter-rater R)

- Diff = Student, Gen = Rater,Time

$$G = \frac{S}{S + SxR/3 + SxT/2 \text{ (+/o T, R)} + err/6}$$

$$= \frac{3.73}{0.945} =$$

$$3.73 + 0/2 + .061/3 + 1.15/6$$

The MMI Study (admissions OSCE)

Sub	Case A		Case B		Case C			Case F	
	R1	R2	R3	R4	R5	R6	….	R11	R12
1	i1 i2 i3	i1 i2 i3							
2									
.									
.									
.									
18									

The Facets

- Subjects (S)
- Raters (R)
- Items (I)
- Cases (C)

Sources of Variance MMI study

- Student (18 levels)
- Case (6 levels)
- Rater (2 levels)
- Item (3 levels)
- S x C, S x R, S x I, C x I, R x I *
- S x R x I, S x C x I, S x C x R *
- Error (SxCxRxI)

Strictly speaking Rater is “nested” within case, so some interactions absent. We ignore.

* These terms are not needed for the calculation

Variance components

Student	0.79
S x I	0.0
S x R	0.10
S x C	0.39
SxCxR	0.75
SxCxI	0.12
SxRxI	0.03
Error	0.17

G coefficients -- General Strategy

$$G = \frac{\text{Total variance - Error(Facet of Gen)}}{\text{Total variance}}$$

Variance components

Student	0.79
S x I	0.00
S x R	0.10
S x C	0.39
SxCxR	0.75
SxCxI	0.12
SxRxI	0.03
Error(SRCI)	0.17
TOTAL	**2.35**

Inter-rater Reliability

Var(SxR) = .10

Var(SxRxI)= .03

Var(SxRxC)= 0.75

Var(error) = 0.17

$$G(\text{rater}) = \frac{2.35 - 1.05}{2.35} = 0.55$$

Inter-Item Reliability

Var(SxI) = .00

Var(SxRxI)= .03

Var(SxCxI)= 0.12

Var(error) = 0.17

$$G(item) = \frac{2.35 - 0.32}{2.35} = 0.86$$

Inter-Case Reliability

Var(SxC) = .39

Var(SxCxI)= .12

Var(SxRxC)= 0.75

Var(error) = 0.17

$$G(case) = \frac{2.35 - 1.43}{2.35} = 0.39$$

G theory equivalent of the Spearman - Brown formula

- Often want to determine the reliability of the average across observations (items, raters, cases) -- the total score

- Approach is to simply divide any variance component using that facet by the number of levels of that facet

Example -- the MMI

- What is the generalizability of the average score of the two raters?

  (divide R terms by 2)

- What is the generalizability of the total score of the 3 items?

  (divide I terms by 3)

- What is the generalizability of the average score of the 6 cases, 2 raters, and 3 items ?

  (divide R terms by 2, I terms by 3, C terms by 6
  IR term by 6, IC term by 18, RC term by 12,
  IRC term by 36)

- What is the generalizability of the average score of the two raters?

(divide R terms by 2)

Den =.79 +(.10+.75+.03+.17)/2 + (0+.39+.12) = 1.82

$$G = \frac{1.82 - (1.05/2)}{1.82} = 0.71$$

- What is the generalizability of the total score of the 3 items?

(divide I terms by 3)

Den =.79 +(0+.12+.03+.17)/3 + (.10+.39+.75) = 2.15

$$G = \frac{2.15 - (0.32/3)}{2.15} = 0.94$$

- What is the generalizability of the average score of the 6 cases, 2 raters, and 3 items ?

(divide R terms by 2, I terms by 3, C terms by 6
IR term by 6, IC term by 18, RC term by 12,
IRC term by 36)

Den =.79 +.10/2+.75/12+.03/6 + 0/3+.39/6+.12/24 +.75/18+.17/36

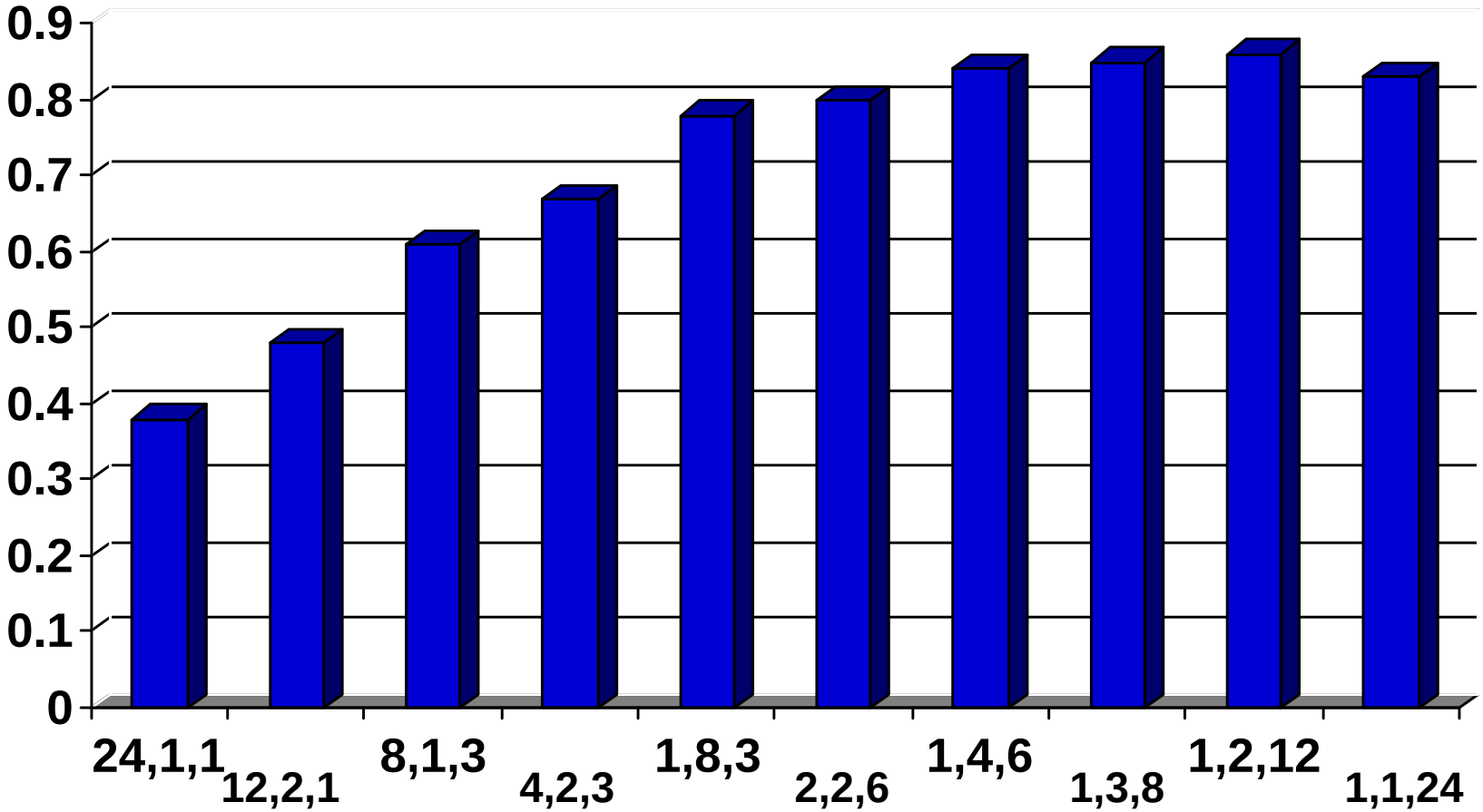
= 0.98

$$G = \frac{0.79}{0.98} = 0.80$$

From G study to D study

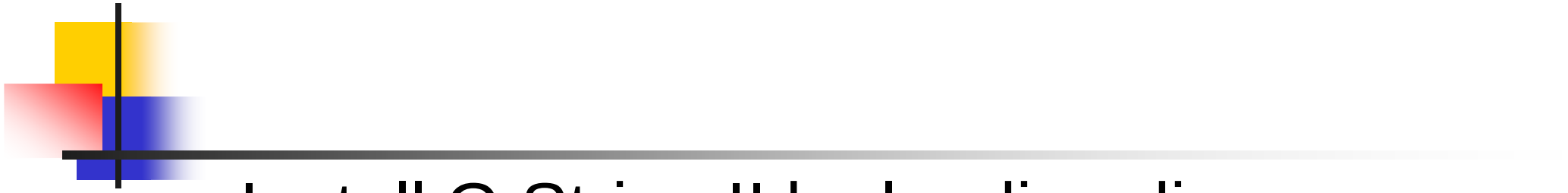
- Up to now, estimating variance components and G coefficients for the original study design --“G study”
- However, with estimated variance components, can now determine:
  - optimum numbers of items, cases, raters for a given total number of observations
  - number of observations of each type to achieve a particular level of reliability
- Decision or “D study”

Overall Reliability by no.items (I), no.raters(r), no.cases (c)



LET’s GO TO WORK

-



- Install G String II by loading disc
  - Click on “Quickie” folder
  - Click on “Setup_G_String.msi
- If you get a message related to dot.net
  - Click on “dotnetfx” and install
  - Click on “Quickie” and proceed as above

- On Desktop
  - Click on “G_STRING_INT.EXE
- In G String:
  - Click on “Start”

Example 2 -- Key Features

- How many questions should we have per case?
  - Depends on variance due to items, cases
  - It takes time to read the case, so fewer questions/case means fewer questions overall

- Assume 2 min. to read case / 2 min to answer each question.
- In 2 hour exam:

	No. Cases	No. Questions
1 q / case	30	30
2 q / case	20	40
3 q / case	15	45
4 q / case	12	48
5 q / case	10	50

$$G = \frac{\text{Var(sub)}}{\text{Var(sub)} + [\text{Var(SxC)} + \text{Var(SxI)}/i\ ]/c}.$$

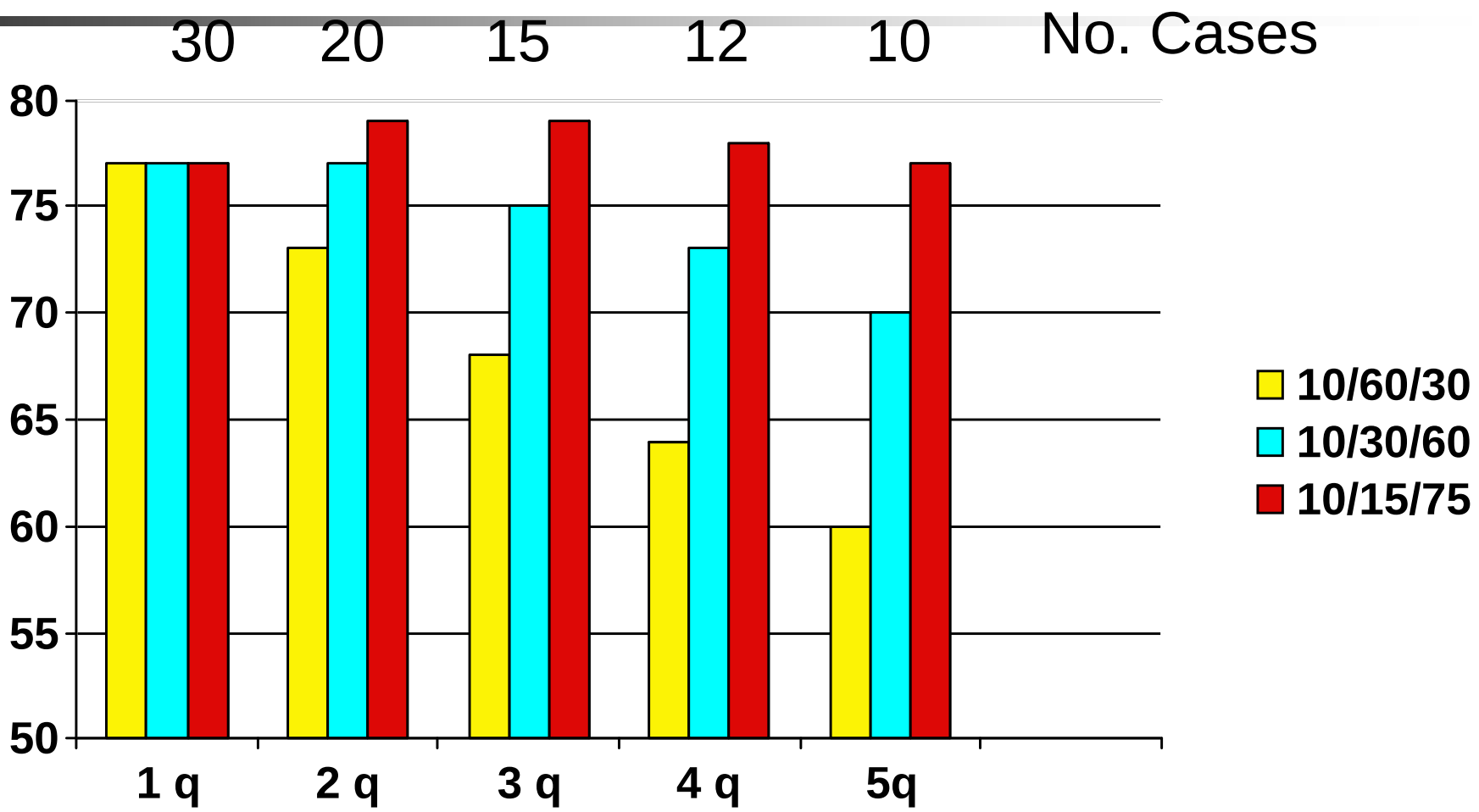
Assume:

Case 1: Var(sub) = 10; Var (SxC)= 60; Var(SxI)=30

Case 2: Var(sub) = 10; Var (SxC)= 30; Var(SxI)=60

Case 3: Var(sub) = 10; Var (SxC)= 15; Var(SxI)=75

Reliability by no. cases/questions



SUMMARY

- G THEORY is an extension of CTT
  - Recognizes multiple sources (facets) of variance
  - All included in a single study design
  - Different G coefficients depending on facets of:
    - Generalization
    - Differentiation
    - Fixed
  - D study to devise optimal strategy

SUMMARY

- Creating Scales:
  - Use >= 7 points on the scale, >= 4 descriptors
  - Score by simple summing of items
  - Do NOT categorize until it’s over

- Measurement and Statistics
  - The Cronbach conundrum
    - Reliability and responsiveness share an error term
  - Use pretest carefully, ensure it’s appropriate
  - Analyze with ANCOVA, not diff score

The Last Word

- “If you can not measure it, you can not improve it.”

  Lord Kelvin

- “It is no measure of health to be well adjusted to a profoundly sick society.”

  Jiddu Krishnamurti

A Brief Commercial Message