Lesson 4
Evaluation of the measurement instrument:
items analysis

1

1. Introduction

- Items can adopt different formats and assess:
  - cognitive variables (skills, performance, etc.) where there are right and wrong answers.
  - non-cognitive variables (attitudes, interests, values, etc.) where there are not right and wrong answers.
- The statistics that we present are used primarily with skills or performance items.

2

1. Introduction

- To carry out the analysis of the items, it should be available:
  - A data matrix with the participants' responses to each item.
    - To analyze test scores and the responses to the correct alternative, the matrix will take the form of ones (right answers) and zeros (wrong answers).
    - To analyze incorrect alternatives, it should appear specific options selected by each participant in the matrix.
- The analysis to carry out are:
  - Difficulty
  - Discrimination
  - Reliability
  - Validity
  - Distractors
  - Differential item functioning

3

1. Introduction

- Empirical difficulty of an item: proportion of participants who answer it correctly.
- Discriminative power: the ability of the item to distinguish the participants with different level in the trait measured.
- Both statistics are directly related to the mean and variance of total test scores.
- The reliability and validity of the items are related to the standard deviation of the test, and indicate the possible contribution of each item to the reliability and validity of the total scores of the test.

4

2. Item difficulty

- Proportion of participants who answered the item correctly:
  - One of the most popular indices to quantify the difficulty of the dichotomous or dichotomized items.
- The difficulty is considered a relative index because it depends on:
  - Number of people who try to answer the item.
  - Their characteristics (e.g., if they are more or less prepared to do the test).

$$ID = \frac{R}{N}$$

R: number of right answers.

N: number of participants that answered the item.

- It ranges between 0 and 1.
  - 0: No one answered the item correctly. It is extremely difficult.
  - 1: All the participants answered correctly the item. It is extremely easy.

5

2. Item difficulty

- Example: A performance item in math is answered by 10 participants. The results are presented in the table below:

Participant	a	b	c	d	e	f	g	h	i	j
Answer	1	1	1	1	0	1	0	1	1	0

Calculate the item difficulty.

6

2. Item difficulty

$$ID = \frac{7}{10} = 0.70$$

- The obtained value does not indicate whether the item is good or bad. It represents how hard it has been for the sample of participants who tried to answer it.
- It can be considered an easy item.

7

2. Item difficulty

- The ID is directly related to the mean and variance of the test. In dichotomous items:

$$ID = \frac{\sum_{j=1}^{n} X_j}{N}$$

$X_j =$ 1 or 0 according to success or failure in the item

$\sum_{j=1}^{n} X_j =$ the number of correct answers

- The sum of all the scores obtained by the participants that answered this item is equal to the number of correct answers. Therefore, the item difficulty is equal to its mean.
- If we generalize to the total test, the average of the test scores is equal to the sum of all the item difficulties.

8

2. Item difficulty

- The relationship between the difficulty and the variance of the test is also direct. In dichotomous items:

$$S_j^2 = p_j q_j$$
$$p_j = ID$$
$$q_j = 1 - p_j$$

  - p_j is the proportion of participants that answered correctly.
  - Maximum variance is achieved by an item when $p_j = 0.5$
  - An item is appropriate when it is answered by different participants and causes in them different answers.

9

2. Item difficulty
2.1. Correction of right answers by chance

- The fact of answer an item correctly depends not only on the participants' knowledge, but also on participants' luck when they do not know the answer.
- The higher the number of distractors is, the less probable is to answer correctly at random.
- It is advisable to correct the ID:

$$ID_C = \frac{R}{N} - \frac{\frac{W}{K-1}}{N} = p - \frac{q}{K-1}$$

(negative values can be found)

ID_C=corrected item difficulty

R = right answers

W = wrong answers

p = proportion of right answers

q = proportion of wrong answers

K = number of alternatives

N = number of participants that answered the item

p + q = 1

10

2. Item difficulty
2.1. Correction of right answers by chance

Participants	Item 1	Item 2	Item 3	Item 4	Item 5
A	1	1	1	1	1
B	1	0	1	0	1
C	1	1	0	1	0
D	1	0	0	1	0
E	0	1	0	1	1
F	1	0	0	1	0
G	0	1	1	1	0
H	1	0	0	1	0
I	1	1	0	0	0
J	0	0	0	1	1

Example. Test composed by items with 3 alternatives. Calculate ID and IDc for each item.

11

2. Item difficulty
2.1. Correction of right answers by chance

Participants	Item 1	Item 2	Item 3	Item 4	Item 5
A	1	1	1	1	1
B	1	0	1	0	1
C	1	1	0	1	0
D	1	0	0	1	0
E	0	1	0	1	1
F	1	0	0	1	0
G	0	1	1	1	0
H	1	0	0	1	0
I	1	1	0	0	0
J	0	0	0	1	1
R					
W					
ID					
IDc					

12

Participants	Item 1	Item 2	Item 3	Item 4	Item 5
A	1	1	1	1	1
B	1	0	1	0	1
C	1	1	0	1	0
D	1	0	0	1	0
E	0	1	0	1	1
F	1	0	0	1	0
G	0	1	1	1	0
H	1	0	0	1	0
I	1	1	0	0	0
J	0	0	0	1	1
R	**7**	**5**	**3**	**8**	**4**
W	**3**	**5**	**7**	**2**	**6**
ID	**0.7**	**0.5**	**0.3**	**0.8**	**0.4**
IDc	**0.55**	**0.25**	**-0.05**	**0.7**	**0.1**

When the items are more difficult, the correction is higher.

13

2. Item difficulty
2.1. Correction of right answers by chance
Recommendations

- The items with extreme index of difficulty are eliminated from the final test.
- We will get better psychometric results when the majority of the items has a medium difficulty.
- Easy items should be included, preferably at the beginning, to measure the less competent participants.
- Difficult items should also be included to measure the most competent participants.

14

2. Item difficulty
2.2. Comparison between items

To determine if the degree of difficulty of two items is equivalent, we can calculate the X^2 (Harris y Pearlman 1977):

	Item 1	
Item 2	R	W
R	a	b
W	c	d

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c}$$

$X^2 \leq X^2_{(\alpha, 1)}$: The null hypothesis is accepted. The degree of difficulty of both items is equivalent.

$X^2 > X^2_{(\alpha, 1)}$: The null hypothesis is rejected. The degree of difficulty of both items is statistically different.

15

2. Item difficulty
2.2. Comparison between items

200 participants answered two items:

	Item 1	
Item 2	Right answer	Wrong answer
Right answer	65 (a)	35 (b)
Wrong answer	35 (c)	65 (d)

¿Is the degree of difficulty equivalent in the two items? (LC = 95%).

16

2. Item difficulty
2.2. Comparison between items

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} = \frac{(|35-35|-1)^2}{70} = 0.014$$

$$\chi^2_{(\alpha,1)} = \chi^2_{(0.05,1)} = 3.84$$

0.014 < 3.84: The null hypothesis is accepted. The degree of difficulty of both items is equivalent.

17

3. Discrimination

- The participants with higher scores in the test should obtain a higher proportion of right answers in an individual item.
- If an item is not useful to differentiate between participants based on their skill level, it should be deleted.

18

3. Discrimination

3.1. Item discrimination index based on extreme groups (D)

- It is based on the proportions of right answers in the extreme groups of ability (upper and lower 25 or 27% of the total sample).
  - The upper 25 or 27% are the participants who obtained higher scores than the 75 or 73% of the sample (they are in percentile 75 or 73, or over it).
- After forming the groups, we calculate:

$$D = p_u - p_l$$

  - p_u = proportion of right answers in the upper group.
  - p_l = proportion of right answers in the lower group.

19

3. Discrimination

3.1. Item discrimination index based on extreme groups (D)

- D index ranges between -1 and 1.
  - 1= when all the people in the upper group answered the item correctly and all the people in the lower group answered it incorrectly.
  - 0= item is equally answered correctly in both groups.
  - Negative values= the less competent participants answered the item correctly in more occasions than the most competent ones (the item confused the most skilled participants).

20

3. Discrimination
3.1. Item discrimination index based on extreme groups (D)

Interpretation of D values (Ebel, 1965)

Values	Interpretation
$D \geq 0.40$	The item discriminates very well
$0.30 \leq D \leq 0.39$	The item discriminates well
$0.20 \leq D \leq 0.29$	The item discriminates slightly
$0.10 \leq D \leq 0.19$	The item needs revision
$D < 0.10$	The item is useless

21

3. Discrimination
3.1. Item discrimination index based on extreme groups (D)

Example. The table below presents the answers given by 370 participants in an item with 3 alternatives (A, B, C), where B is the correct option.
The rows present the frequency of participants who selected each alternative and obtained scores over and under the 27% of their sample in the total test, and the group formed by the central 46%.

	A	B*	C
Upper 27%	19	53	28
Intermediate 46%	52	70	48
Lower 27%	65	19	16

Calculate the corrected difficulty and the discrimination index. Is it an easy item? Does it discriminate well?

22

3. Discrimination

3.1. Item discrimination index based on extreme groups (D)

$$ID_c = p - \frac{q}{k-1} = 0.38 - \frac{0.62}{3-1} = 0.07$$

$$p = \frac{53+70+19}{370} = 0.38$$

$$q = \frac{19+52+65+28+48+16}{370} = 1 - 0.38 = 0.62$$

$$D = p_u - p_l = \frac{53}{19+53+28} - \frac{19}{65+19+16} = \frac{53-19}{100} = 0.34$$

The item is difficult (value close to 0) and discriminates well.

23

3. Discrimination

3.2. Item discrimination indices based on the correlation

- If an item discriminates adequately, the correlation between the scores obtained by participants in that item and the ones obtained in the total test will be positive.
  - participants who obtain high scores in the test are more likely to answer the item correctly.
- Definition: correlation between participants' scores in the item and their scores in the test (Muñiz, 2003).
- The total score of the participants in the test will be calculated discounting the item score.

24

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.1. Correlation coefficient Φ

- When **test scores and item scores** are strictly **dichotomous**.
- It allows to estimate the discrimination of an item with some criterion of interest (e.g., fit and unfit, gender, etc.).
- First, we have to sort data in a 2x2 contingency table.
  - 1= item answered correctly/criterion exceeded.
  - 0= item answered incorrectly/criterion not exceeded.

25

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.1. Correlation coefficient *Φ*

		Item (X)	
		1	**0**
Criterion (Y)	**Fit**	a	b
	Not fit	c	d

$$\phi = \frac{p_{xy} - p_x p_y}{\sqrt{p_x q_x p_y q_y}}$$

$$p_{xy} = \frac{a}{N} \qquad p_x = \frac{(a+c)}{N} \qquad q_x = \frac{(b+d)}{N}$$

$$p_y = \frac{(a+b)}{N} \qquad q_y = \frac{(c+d)}{N}$$

26

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.1. Correlation coefficient Φ

Example. The following table shows the sorted results from 50 participants who did the last psychometrics exam.

		Item 5 (X)	
		1	**0**
Criterion (Y)	**Fit**	30 (a)	5 (b)
	Not fit	5 (c)	10 (d)

Calculate the correlation coefficient φ

27

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.1. Correlation coefficient Φ

		Item 5 (X)		
		1	**0**	
Criterion (Y)	**Fit**	p_{xy} 30/50=0.6	5	p_y 35/50=0.7
	Not fit	5	10	q_y 15/50=0.3
		p_x 35/50=0.7	q_x 15/30=0.3	N=50

28

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.1. Correlation coefficient Φ

$$\phi = \frac{p_{xy} - p_x p_y}{\sqrt{p_x q_x p_y q_y}} = \frac{0.6 - 0.7 * 0.7}{\sqrt{0.7 * 0.3 * 0.7 * 0.3}} = 0.52$$

- There is a high correlation between the item and the criterion. That is, those participants who answered correctly the item usually passed the psychometrics exam.

29

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.2. Point-biserial correlation

- When the **item** is a **dichotomous** variable and the **test** score is **continuous**.

$$r_{pb} = \frac{\overline{X}_1 - \overline{X}_T}{S_X} \sqrt{\frac{p}{q}}$$

$\overline{X}_1$ = Mean in the test of participants that answered the item correctly.
$\overline{X}_T$ = Mean of the test.
S_X = Standard deviation of the test.
p = Proportion of participants that answered the item correctly.
q = proportion of participants that answered the item incorrectly.

- Remove the item score from the test score.

30

3. Discrimination
3.2 Item discrimination index based on the correlation
3.2.2. Point-biserial correlation

Example. The following table shows the responses of 5 participants to 4 items. Calculate the point-biserial correlation of the second item.

	Items			
Participants	**1**	**2**	**3**	**4**
A	0	1	0	1
B	1	1	0	1
C	1	1	1	1
D	0	0	0	1
E	1	1	1	0

31

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.2. Point-biserial correlation

	Items				Total		
Participants	**1**	**2**	**3**	**4**	**X**	**(X-i)**	**$(X-i)^2$**
A	0	1	0	1	2	1	1
B	1	1	0	1	3	2	4
C	1	1	1	1	4	3	9
D	0	0	0	1	1	1	1
E	1	1	1	0	3	2	4
∑						**9**	**19**

32

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.2. Point-biserial correlation

- Participants who answered correctly the item are A, B, C and E; so their mean is:

$$\overline{X}_1 = \frac{1+2+3+2}{4} = 2$$

- The total mean is:

$$\overline{X}_T = \frac{9}{5} = 1.8$$

- The standard deviation of the test is:

$$S_X = \sqrt{\frac{\sum X^2}{N} - \overline{X}^2} = \sqrt{\frac{19}{5} - 1.8^2} = \sqrt{0.56} = 0.75$$

33

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.2. Point-biserial correlation

$$p = \frac{4}{5} = 0.8$$

$$q = \frac{1}{5} = 0.2$$

$$r_{pb} = \frac{\overline{X}_1 - \overline{X}_T}{S_X}\sqrt{\frac{p}{q}} = \frac{2-1.8}{0.75}\sqrt{\frac{0.8}{0.2}} = 0.54$$

34

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.3. Biserial correlation

- When both **item and test** score are inherently **continuous variables**, although one is dichotomized (the item).

$$r_b = \frac{\overline{X}_1 - \overline{X}_T}{S_X}\frac{p}{y}$$

y = height in the normal curve corresponding to the typical score that leaves beneath a probability equal to p (see table).

- We can find values greater than 1, especially when one of the variables is not normal.

- **Example**. Based on the table of the previous example, calculate the biserial correlation of item 3.

35

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.3. Biserial correlation

	Items				Total		
Participants	**1**	**2**	**3**	**4**	**X**	**(X-i)**	**$(X-i)^2$**
A	0	1	0	1	2	2	4
B	1	1	0	1	3	3	9
C	1	1	1	1	4	3	9
D	0	0	0	1	1	1	1
E	1	1	1	0	3	2	4
∑						**11**	**27**

36

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.3. Biserial correlation

- Participants who answered correctly the item are C and E; so their mean is:

$$\overline{X}_1 = \frac{3+2}{2} = 2.5$$

- The total mean is:

$$\overline{X}_T = \frac{11}{5} = 2.2$$

- The standard deviation of the test is:

$$S_X = \sqrt{\frac{\sum X^2}{N} - \overline{X}^2} = \sqrt{\frac{27}{5} - 2.2^2} = \sqrt{5.4 - 4.84} = \sqrt{0.56} = 0.75$$

37

3. Discrimination
3.2. Item discrimination index based on the correlation
3.2.3. Biserial correlation

$$p = \frac{2}{5} = 0.4$$

$$r_b = \frac{\overline{X}_1 - \overline{X}_T}{S_X} \frac{p}{y} = \frac{2.5 - 2.2}{0.75} \frac{0.4}{0.3863} = 0.4 * 1.03 = 0.41$$

Because the value *p*=0.4 does not appear in the first column of the table, we look its complement up (0.6), which is associated with an *y*=0.3863.

38

3. Discrimination
3.3. Discrimination in attitude items

- There are no right or wrong answers but the participant must be placed in the continuum established based on the degree of the measured attribute.
- Correlation between item scores and test scores.
  - Because items are not dichotomous, Pearson correlation coefficient is used.
    - That coefficient can be interpreted as a Homogeneity Index (HI). It indicates how much the item is measuring the same dimension or attitude as the rest of the items of the scale.

39

3. Discrimination
3.3. Discrimination in attitude items

$$R_{jX} = \frac{N\sum JX - \sum J \sum X}{\sqrt{\left[N\sum J^2 - (\sum J)^2\right]\left[N\sum X^2 - (\sum X)^2\right]}}$$

- N = sample size
- ∑J = sum of the scores in the item J
- ∑X = sum of the scores in the scale
- R_{jx} = correlation between the scores obtained in the item J and the scale

- Items with a HI < 0.2 should be eliminated.
- Correction: subtract the total score minus the item score in each participant or apply the formula below:

$$R_{j(x-j)} = \frac{R_{jx} S_x - S_j}{\sqrt{S_x^2 + S_j^2 - 2 R_{jx} S_x S_j}}$$

40

3. Discrimination
3.3. Discrimination in attitude items

Example. The table below presents the answers of 5 people to 4 attitudes items. Calculate the discrimination of item 4 by Pearson correlation.

	Items			
participants	**X1**	**X2**	**X3**	**X4**
A	2	4	4	3
B	3	4	3	5
C	5	2	4	3
D	3	5	2	4
E	4	5	2	5

41

3. Discrimination
3.3. Discrimination in attitude items

	Items				X_T	X_4X_T	X^2_4	X^2_T
participants	**X1**	**X2**	**X3**	**X4**				
A	2	4	4	3	13	39	9	169
B	3	4	3	5	15	75	25	225
C	5	2	4	3	14	42	9	196
D	3	5	2	4	14	56	16	196
E	4	5	2	5	16	80	25	256
∑				**20**	**72**	**292**	**84**	**1042**

42

3. Discrimination
3.3. Discrimination in attitude items

- The correlation or IH between item 4 and total score of the test will be:

$$R_{jx} = \frac{N\sum JX - \sum J \sum X}{\sqrt{[N\sum J^2 - (\sum J)^2][N\sum X^2 - (\sum X)^2]}} = \frac{5*292 - 20*72}{\sqrt{[5*84 - 20^2][5*1042 - 72^2]}} = 0.88$$

- Inflated result because item 4 score is included in total score. Correction:

$$R_{j(x-j)} = \frac{R_{jx}S_x - S_j}{\sqrt{S_x^2 + S_j^2 - 2R_{jx}S_xS_j}} = \frac{0.88*1.02 - 0.89}{\sqrt{1.04 + 0.80 - 2*0.88*1.02*0.89}} = 0.01$$

$$\overline{X}_4 = \frac{\sum X_4}{N} = \frac{20}{5} = 4 \qquad S_{X_4} = \sqrt{\frac{\sum X_4^2}{N} - \overline{X}_4^2} = \sqrt{\frac{84}{5} - 4^2} = \sqrt{0.8} = 0.89$$

$$\overline{X}_T = \frac{\sum X_T}{N} = \frac{72}{5} = 14.4 \qquad S_{X_T} = \sqrt{\frac{\sum X_T^2}{N} - \overline{X}_T^2} = \sqrt{\frac{1042}{5} - 14.4^2} = \sqrt{1.04} = 1.02$$

43

3. Discrimination
3.3. Discrimination in attitude items

- The big difference when applying the correction is due to the small number of items that we have used in the example.
  - As the number of items increases, that effect decreases because the influence of item scores on the total score is getting smaller. With more than 25 items, the result is very close.

44

3. Discrimination
3.3. Discrimination in attitude items

- Another procedure:
  - Useful but less efficient than the previous because it does not use the entire sample.
  - Determine whether the item mean for the participants with higher scores on the total test is statistically higher than the mean of those with lower scores. It is common to use 25% or 27% of participants with best and worst scores.
  - Once the groups are identified, we calculate if the mean difference is statistically significant by Student T test.
  - Ho: means in upper group is equal or smaller than in the low group

$$T = \frac{\overline{X}_{uj} - \overline{X}_{lj}}{\sqrt{\frac{(n_u - 1)S_{uj}^2 + (n_l - 1)S_{lj}^2}{n_u + n_l - 2}\left[\frac{1}{n_u} + \frac{1}{n_l}\right]}}$$

45

3. Discrimination
3.3. Discrimination in attitude items

$\overline{X}_{uj}$ = mean of the scores obtained in the item by the 25% of the participants that obtained the highest scores in the test.

$\overline{X}_{lj}$ = mean of the scores obtained in the item by the 25% of the participants that obtained the lowest scores in the test.

S_{uj}^2 = variance of the scores obtained in the item by the 25% of the participants that obtained the highest scores in the test.

S_{lj}^2 = variance of the scores obtained in the item by the 25% of the participants that obtained the lowest scores in the test.

n_u and n_l = number of participants in the upper and the lower group respectively.

- Conclusions:
  - $T \leq T_{(\alpha,nu+nl-2)}$ – Null hypothesis is accepted. There are not statistical differences between means. The item does not discriminate adequately.
  - $T > T_{(\alpha,nu+nl-2)}$ – Null hypothesis is rejected. There are statistical differences between means. The item discriminates adequately.
- Student T test is used when the scores in the item and the scale are distributed normally, and their variances are equal. If some of these assumptions are violated, a non-parametric test should be used (e.g., Mann-Whitney U).

46

3. Discrimination
3.3. Discrimination in attitude items

Exercise: using the data presented in the last example, calculate Student T test for item 2 (α=0.05).

- To calculate the discrimination of item 2 by Student T Test, we have to do groups with extreme scores. Because of didactic reasons, we are going to use just 2 participants to form those groups.

	Participants	X_2
Upper group	E (16)	5
	B (15)	4
	Participants	**X_2**
Lower group	A (13)	4
	C (14)	2

47

3. Discrimination
3.3. Discrimination in attitude items

	Participants	X_2	X_2^2
Upper group	E (16)	5	25
	B (15)	4	16
	∑	9	41
	Participants	**X_2**	**X_2^2**
Lower group	A (13)	4	16
	C (14)	2	4
	∑	6	20

$$\overline{X}_{uj} = \frac{\sum X_{uj}}{n_u} = \frac{9}{2} = 4.5$$

$$\overline{X}_{lj} = \frac{\sum X_{lj}}{n_l} = \frac{6}{2} = 3$$

48

3. Discrimination
3.3. Discrimination in attitude items

$$S_{uj}^2 = \frac{\sum X_{uj}^2}{n_u} - \overline{X}_{uj}^2 = \frac{41}{2} - 4.5^2 = 20.5 - 20.25 = 0.25$$

$$S_{lj}^2 = \frac{\sum X_{lj}^2}{n_l} - \overline{X}_{lj}^2 = \frac{20}{2} - 3^2 = 10 - 9 = 1$$

$$T = \frac{\overline{X}_{uj} - \overline{X}_{lj}}{\sqrt{\frac{(n_u - 1)S_{uj}^2 + (n_l - 1)S_{lj}^2}{n_u + n_l - 2}\left[\frac{1}{n_u} + \frac{1}{n_l}\right]}} = \frac{4.5 - 3}{\frac{(2-1)0.25 + (2-1)1}{2+2-2}\left(\frac{1}{2} + \frac{1}{2}\right)} = 1.9$$

One tail: $T_{(\alpha, nu+nl-2)} = T_{(0.05,\ 2+2-2)} = T_{(0.05,\ 2)} = 2.92$

1.9 < 2.92 – The null hypothesis is accepted. The upper group doesn't present higher significant statistical mean difference with respect the lower group. The item does not discriminate adequately.

49

3. Discrimination
3.4. Factors that affect the discrimination
3.4.1. Variability

- Relation between test variability and item discrimination:

$$S_x = \sum_{j=1}^{n} S_j r_{jx}$$

S_x = Standard deviation of the test

S_j = Standard deviation of the item

r_{jx} = Discrimination index of item j

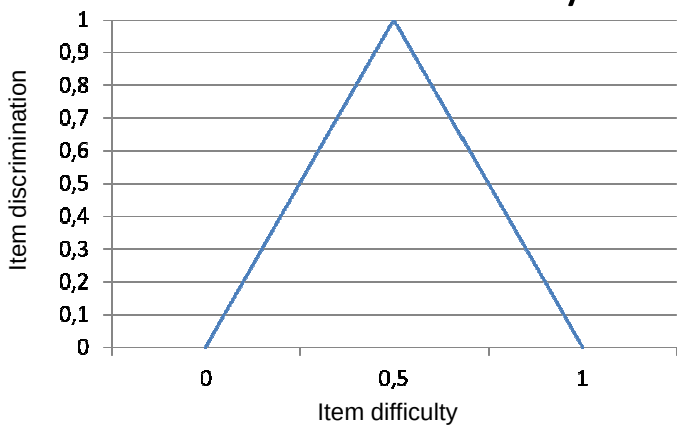
- If the test is composed by dichotomous items:

$$S_x^2 = \sum_{j=1}^{n} p_j q_j r_{jx}^2; S_x = \sqrt{\sum_{j=1}^{n} p_j q_j r_{jx}^2}$$

- To maximize the discriminative ability of one test, we have to consider together both the difficulty (p_j) and the discrimination (r_{jx}) of its items.
  - It is achieved when discrimination is maximum (r_{jx}=1) and the difficulty is medium (p_j=0.5).

50

3. Discrimination
3.4. Factors that affect the discrimination
3.4.2. Item difficulty



An item reaches its maximum discriminative power when its difficulty is medium.

51

3. Discrimination
3.4. Factors that affect the discrimination
3.4.3. Dimensionality of the test

- When we are constructing a test, usually we try to measure one single construct (one-dimensionality).
- In multidimensional tests, item discrimination should be estimated considering only the items that are associated with each dimension.

52

3. Discrimination
3.4. Factors that affect the discrimination
3.4.4. Test reliability

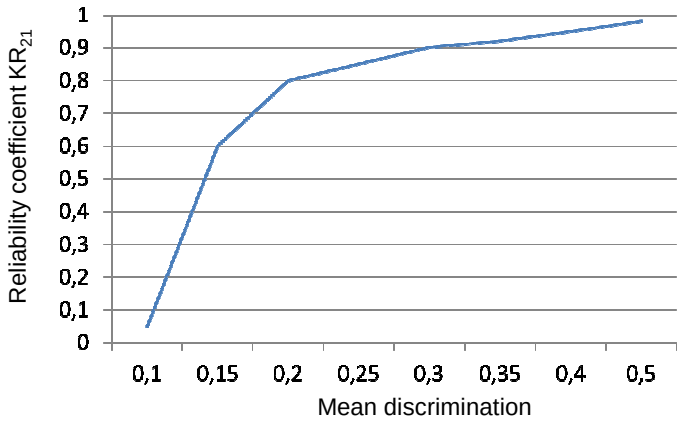
- If discrimination is defined as the correlation between scores obtained by participants in the item and the test, then reliability and discrimination are closely related.
- It is possible to express the Cronbach alpha coefficient from the discrimination of items:

$$\alpha = \frac{n}{n-1}\left(1-\frac{\sum_{j=1}^{n} S_j^2}{S_x^2}\right) = \frac{n}{n-1}\left(1-\frac{\sum_{j=1}^{n} S_j^2}{\left[\sum_{j=1}^{n} S_j r_{jx}\right]^2}\right)$$

- Small values in item discrimination are typically associated with unreliable tests.

53

3. Discrimination
3.4. Factors that affect the discrimination
3.4.4. Test reliability



As the mean discrimination of the test increases, so does the reliability coefficient.

54

4. Indices of reliability and validity of the items
4.1. Reliability index

- To quantify the degree in which an item is measuring accurately the attribute of interest.

$$RI = S_j D_j$$

S_j = Standard deviation of the scores in the item.

D_j = Discrimination index of the item.

- When any correlation coefficient is used to calculate the discrimination of items,

$$RI = S_j r_{jx}$$

$$S_j^2 = pq \qquad \sum RI^2 = S_X^2$$

55

4. Indices of reliability and validity of the items
4.1. Reliability index

- To the extent that we select items with higher RI, the better the reliability of the test will be.
- Highest possible value of RI = 1.

- Example: Having the information presented in the table below, calculate the RI of item 4.

	p	r_{bp}
Item 4	0.47	0.5

56

4. Indices of reliability and validity of the items
4.1. Reliability index

$$RI = S_j r_{jx} = 0.5 * 0.5 = 0.25$$

$$S_j^2 = pq = 0.47 * 0.53 = 0.25$$

$$q = 1 - p = 1 - 0.47 = 0.53$$

$$S_j = \sqrt{S_j^2} = \sqrt{0.25} = 0.5$$

57

4. Indices of reliability and validity of the items
4.2. Validity index

- The validity of an item involves the correlation of the scores obtained by a sample of participants in the item with the scores obtained by the same participants in any external criterion of our interest.
  - It serves to determine the degree in which each item of one test contributes successfully to make predictions about that external criterion.

$$VI = r_{jy}$$

- In the case that the criterion is a continuous variable and the item is a dichotomous variable, we are going to use the point-biserial correlation; but it is not necessary to subtract from the total score of the external criterion the item score because it is not included.

$$VI = r_{pbjy}$$

58

4. Indices of reliability and validity of the items
4.2. Validity index

- Test validity (r_{xy}) can be expressed in connection with the VI of the items. The higher VI of the items are, the more optimized the validity of the test will be.

$$r_{xy} = \frac{\sum S_j r_{jy}}{\sum S_j r_{jx}} = \frac{\sum VI}{\sum RI}$$

- This formula allows us to see how the validity of the test can be estimated from the discrimination index of each item (r_{jx}), their validity indexes (r_{jy}) and their difficulty indexes ($S_j^2 = p_j q_j$).

59

4. Indices of reliability and validity of the items
4.2. Validity index

- Paradox in the selection of items: if we want to select items to maximize the reliability of the test we have to choose those items with a high discrimination index (r_{jx}), but this would lead us to reduce the validity of the test (r_{xy}) because it increases when validity indexes (VI) are high and reliability indexes (RI) are low.

60

4. Indices of reliability and validity of the items
4.2. Validity index

Example. The table below presents the scores of 5 participants in a test with 3 items.

Participants	Item 1	Item 2	Item 3
A	0	0	1
B	1	1	1
C	1	0	0
D	1	1	1
E	1	1	1
r_{jy}	0.2	0.4	0.6

Calculate the validity index of the test (r_{xy}).

61

4. Indices of reliability and validity of the items
4.2. Validity index

$$r_{xy} = \frac{\sum S_j r_{jy}}{\sum S_j r_{jx}} = \frac{0.4*0.2 + 0.49*0.4 + 0.4*0.6}{0.4*0.25 + 0.49*0.99 + 0.4*0.25} = \frac{0.516}{0.685} = 0.75$$

$$S_j^2 = p_j q_j \Rightarrow S_j = \sqrt{S_j^2}$$

$$S_1^2 = \frac{4}{5} * \frac{1}{5} = 0.8*0.2 = 0.16 \Rightarrow S_1 = \sqrt{0.16} = 0.4$$

$$S_2^2 = \frac{3}{5} * \frac{2}{5} = 0.6*0.4 = 0.24 \Rightarrow S_2 = \sqrt{0.24} = 0.49$$

$$S_3^2 = \frac{4}{5} * \frac{1}{5} = 0.8*0.2 = 0.16 \Rightarrow S_3 = \sqrt{0.16} = 0.4$$

62

4. Indices of reliability and validity of the items
4.2. Validity index

	1	2	3	X	(X-it1)	(X-it2)	(X-it3)	$(X\text{-it1})^2$	$(X\text{-it2})^2$	$(X\text{-it3})^2$
A	0	0	1	1	1	1	0	1	1	0
B	1	1	1	3	2	2	2	4	4	4
C	1	0	0	1	0	1	1	0	1	1
D	1	1	1	3	2	2	2	4	4	4
E	1	1	1	3	2	2	2	4	4	4
r_{jy}	0.2	0.4	0.6		Σ=7	Σ=8	Σ=7	Σ=13	Σ=14	Σ=13

63

4. Indices of reliability and validity of the items
4.2. Validity index

$$r_{pb_1} = \frac{\overline{X}_1 - \overline{X}_T}{S_X}\sqrt{\frac{p}{q}} = \frac{1.5-1.4}{0.8}\sqrt{\frac{0.8}{0.2}} = \frac{0.1}{0.8}\sqrt{4} = 0.125*2 = 0.25$$

$$\overline{X}_1 = \frac{2+0+2+2}{4} = \frac{6}{4} = 1.5$$

$$\overline{X}_T = \frac{7}{5} = 1.4$$

$$S_X = \sqrt{\frac{13}{5} - 1.4^2} = \sqrt{2.6-1.96} = \sqrt{0.64} = 0.8$$

$$p = \frac{4}{5} = 0.8$$

$$q = 1 - p = 1 - 0.8 = 0.2$$

64

4. Indices of reliability and validity of the items
4.2. Validity index

$$r_{pb_2} = \frac{\overline{X}_1 - \overline{X}_T}{S_X}\sqrt{\frac{p}{q}} = \frac{2-1.6}{0.49}\sqrt{\frac{0.6}{0.4}} = \frac{0.4}{0.49}\sqrt{1.5} = 0.99$$

$$\overline{X}_1 = \frac{2+2+2}{3} = 2$$

$$\overline{X}_T = \frac{8}{5} = 1.6$$

$$S_X = \sqrt{\frac{14}{5} - 1.6^2} = \sqrt{2.8 - 2.56} = \sqrt{0.24} = 0.49$$

$$p = \frac{3}{5} = 0.6$$

$$q = 1 - p = 1 - 0.6 = 0.4$$

65

4. Indices of reliability and validity of the items
4.2. Validity index

$$r_{pb_3} = \frac{\overline{X}_1 - \overline{X}_T}{S_X}\sqrt{\frac{p}{q}} = \frac{1.5-1.4}{0.8}\sqrt{\frac{0.8}{0.2}} = \frac{0.1}{0.8}\sqrt{4} = 0.125 * 2 = 0.25$$

$$\overline{X}_1 = \frac{0+2+2+2}{4} = \frac{6}{4} = 1.5$$

$$\overline{X}_T = \frac{7}{5} = 1.4$$

$$S_X = \sqrt{\frac{13}{5} - 1.4^2} = \sqrt{2.6 - 1.96} = \sqrt{0.64} = 0.8$$

$$p = \frac{4}{5} = 0.8$$

$$q = 1 - p = 1 - 0.8 = 0.2$$

66

5. Analysis of distractors

- It involves investigating in the distribution of participants across the wrong alternatives (distractors), in order to detect possible reasons for the low discrimination of any item or see that some alternatives are not selected by anyone, for example.
- In this analysis, the first step implies:
  - To check that all the incorrect options are chosen by a minimum number of participants. If possible, they should be equally probable.
    - Criteria: each distractor have to be selected by at least the 10% of the sample and there is not many difference between them.
  - That performance on the test of participants who have selected each incorrect alternative is less than the performance of participants that have selected the correct one.
  - It is expected that as the skill level of participants increases, the percentage of those who select incorrect alternatives decrease and vice versa.

67

5. Analysis of distractors
5.1. Same probability of distractors

- Distractors are equally probable if they are selected by a minimum of participants and if they are equally attractive to those who do not know the correct answer.

- X^2 Test:

$$\chi^2 = \sum_{j=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

E_i = Expected (theoretical) frequency.

O_i = Observed frequency.

68

5. Analysis of distractors
5.1. Same probability of distractors

- Degrees of freedom: K -1 (K = number of incorrect alternatives).
- Ho: $E_i = O_i$ (in the participants that do not know the correct answer, the election of any distractor is equally attractive).
- Conclusion:
  - $\chi^2_O \leq \chi^2_{(\alpha,k-1)}$ → The null hypothesis is accepted. The distractors are equally attractive.
  - $\chi^2_O > \chi^2_{(\alpha,k-1)}$ → The null hypothesis is rejected. The distractors are not equally attractive.

69

5. Analysis of distractors
5.1. Same probability of distractors

Example. Determine if the incorrect alternatives are equally attractive (α=0.05).

	A	B*	C
Number of answers	136	142	92

70

5. Analysis of distractors
5.1. Same probability of distractors

$$\chi^2 = \sum_{j=1}^{k} \frac{(E_i - O_i)^2}{E_i} = \frac{(114-136)^2 + (114-92)^2}{114} =$$

$$= \frac{22^2 + 22^2}{114} = \frac{484+484}{114} = \frac{968}{114} = 8.49$$

$$E_i = \frac{136+92}{2} = \frac{228}{2} = 114$$

To be equally probable, each distractor should be selected by 114 participants.

71

5. Analysis of distractors
5.1. Same probability of distractors

$$\chi^2_{(\alpha,k-1)} = \chi^2_{(0.05,2-1)} = \chi^2_{(0.05,1)} = 3.84$$

8.49 > 3.84 → The null hypothesis is rejected. Incorrect alternatives are not equally attractive to all participants, although they met the criterion of being selected by a minimum of 10% of the total sample (N).

$$N = 136 + 142 + 92 = 370$$

$$10\% = \frac{370*10}{100} = 37$$

$$136 > 37$$

$$92 > 37$$

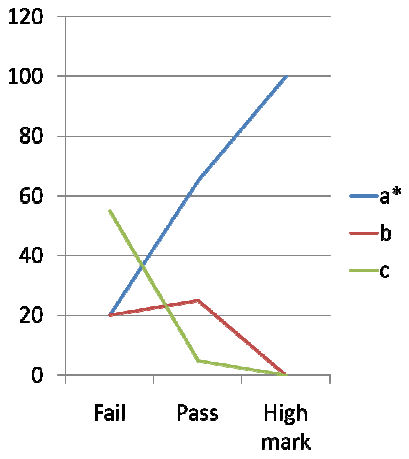
72

5. Analysis of distractors
5.2. Discriminative power of distractors

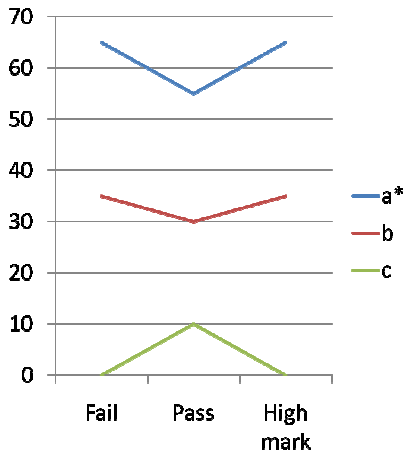
- A distractor is considered good when its correlation with the test scores is negative.
- Correlation is used to quantify the discriminative power of incorrect alternatives. Depending on the kind of variable, we are going to use phi, point-biserial, biserial or Pearson.

73

5. Analysis of distractors
5.2. Discriminative power of distractors



Example of a good item



Example of a bad item

74

5. Analysis of distractors
5.2. Discriminative power of distractors

- Good item:
  - When the mark is getting higher, the correct option (a) is chosen by more participants.
  - When the mark is getting higher, the incorrect options (b and c) are chosen by less participants.
  - Incorrect options (b and c) are equally selected in low marks.
- Bad item:
  - Correct option (a) is equally chosen, regardless of the mark obtained by the participants.
  - Incorrect options (b and c) are also equally chosen, regardless of the mark obtained by the participants.
  - Option c is hardly chosen.

75

5. Analysis of distractors
5.2. Discriminative power of distractors

Example. The table below presents the answers of 5 participants to 4 items. Brackets show the alternatives selected by each participant. The correct alternative is marked with an asterisk. Calculate the discrimination of the distractor b in the item 3.

	Items			
Participants	**1(a*)**	**2(b*)**	**3(a*)**	**4(c*)**
A	0 (b)	1	0 (b)	1
B	1	1	0 (b)	1
C	1	1	1	1
D	0 (c)	0 (a)	0 (b)	1
E	1	1	1	0 (b)

76

5. Analysis of distractors
5.2. Discriminative power of distractors

	Items				Total		
Participants	**1(a*)**	**2(b*)**	**3(a*)**	**4(c*)**	**X**	**(X-i)**	**$(X-i)^2$**
A	0 (b)	1	0 (b)	1	2	2	4
B	1	1	0 (b)	1	3	3	9
C	1	1	1	1	4	3	9
D	0 (c)	0 (a)	0 (b)	1	1	1	1
E	1	1	1	0 (b)	3	2	4
Σ						11	27

$$r_{pb} = \frac{\overline{X}_1 - \overline{X}_T}{S_X}\sqrt{\frac{p}{q}} = \frac{2-2.2}{0.75}\sqrt{\frac{0.4}{0.6}} = -0.22$$

As the result is a negative value, the distractor can be consider good (it was mainly chosen by the participants with lowest level of knowledge).

77

5. Analysis of distractors
5.2. Discriminative power of distractors

Calculations:

- Mean of the test scores of the participants **that selected alternative b** (incorrect) in item 3 (participants A, B and D):

$$\overline{X}_1 = \frac{2+3+1}{3} = \frac{6}{3} = 2$$

- The other calculations are as usual:

$$\overline{X}_T = \frac{11}{5} = 2.2$$

$$S_X = \sqrt{\frac{27}{5} - 2.2^2} = \sqrt{5.4 - 4.84} = \sqrt{0.56} = 0.75$$

$$p = \frac{2}{5} = 0.4$$

$$q = 1 - p = 1 - 0.4 = 0.6$$

78

5. Analysis of distractors
5.2. Discriminative power of distractors

- Visual inspection of the distribution of answers given by a sample to 3-alternative items.

		A	B	C*
Skill level	High	20	25	55
	Low	40	35	25
Statistics	p	0.28	0.5	0.22
	Mean	5	10	9
	r_{pb}	-0.20	0.18	0.29

  - p = proportion of participants that have selected each option.
  - Mean = mean in the test of the participants that selected each alterative.
  - r_{pb} = discrimination index of all the options.

79

5. Analysis of distractors
5.2. Discriminative power of distractors

- Option C (correct answer):
  - Positive discrimination index: the correct alternative is mostly chosen by competent participants.
- Distractor A:
  - Is selected by an acceptable minimum of participants (28%).
  - Is selected by participants less competent in a higher proportion (40 vs. 20 answers; negative discrimination index).
- Distractor B should be revised:
  - Positive discrimination index: it is chosen as correct by the participants with better scores in the test.
  - It has been the most selected (50%).

80

5. Analysis of distractors
5.2. Discriminative power of distractors

In distractors analysis, statistical inference can be used: the mean in the test of participants that choose the correct alternative should be higher than the mean in the test of participants that choose each distractor: ANOVA:

- Independent variable or factor: each item.
  - Conditions: alternatives.
- Dependent variable: the raw score obtained in the test by participants.
- Expected results:
  - There are statistically significant differences between the correct alternative and the incorrect ones.
  - There are not statistically significant differences between incorrect alternatives (same probability).

81

6. Differential item functioning (DIF)

- DIF: procedure to detect biased items.
- Bias: reason why an item benefits some participants across others with the same level of ability, just because they belong to different subpopulations.
- Impact: real differences between groups.
- To sum up, variety obtained in an item could be due to bias or impact. To know it, DIF can be carried out.

82

6. Differential item functioning (DIF)
6.1. Mantel-Haenszel

- One of the most used to detect DIF due to its parsimony.

- Steps:
  1. Detect a variable as a possible cause of the differences.
  2. Form two groups: a reference (RG) and a focal group (FG). The RG is usually the favored one.
  3. Form different levels of aptitude based on the empirical test scores.
  4. Count the number of correct and incorrect answers in each group (RG and FG) and level of ability.

83

6. Differential item functioning (DIF)
6.1. Mantel-Haenszel

	Correct	Incorrect	
RG	A_i	B_i	n_{Ri}
FG	C_i	D_i	n_{Fi}
	n_{1i}	n_{0i}	N_i

$$H_0 : \frac{A_i}{B_i} = \frac{C_i}{D_i}$$ for all the categories

$$\alpha_{MH} = \frac{\sum_{i=1}^{n} \frac{A_i D_i}{N_i}}{\sum_{i=1}^{n} \frac{B_i C_i}{N_i}}$$

84

6. Differential item functioning (DIF)
6.1. Mantel-Haenszel

– Possible results between zero and infinite.

– Interpretation:

- α_{MH} = 1 or close: there is not DIF.
- α_{MH} > 1: there is DIF in favor of the reference group.
- α_{MH} < 1: there is DIF in favor of the focal group.

85

6. Differential item functioning (DIF)
6.1. Mantel-Haenszel

Example. An item of the exam to access to the university is suspected to be damaging to Andalusian students. The results obtained are presented in the table below.

	Non-Andalusian (RG)		Andalusian (FG)	
Exam marks	**Correct**	**Incorrect**	**Correct**	**Incorrect**
0-4	2	7	0	9
5-10	15	51	8	51
11-15	25	48	21	80
16-20	67	14	50	35
21-35	43	8	37	10

Use Mantel-Haenszel method to check if that item presents DIF.

86

6. Differential item functioning (DIF)
6.1. Mantel-Haenszel

Level I of ability (0≤X≤4)			
	Correct	Incorrect	
RG	2	7	
FG	0	9	18

Level II of ability (5≤X≤10)			
	Correct	Incorrect	
RG	15	51	
FG	8	51	125

Level III of ability (11≤X≤15)			
	Correct	Incorrect	
RG	25	48	
FG	21	80	174

Level IV of ability (16≤X≤20)			
	Correct	Incorrect	
RG	67	14	
FG	50	35	166

Level I of ability (21≤X≤35)			
	Correct	Incorrect	
RG	43	8	
FG	37	10	98

87

6. Differential item functioning (DIF)
6.1. Mantel-Haenszel

Aptitude levels	A_iD_i/N_i	B_iC_i/N_i
Level I	2*9/18=1	7*0/18=0
Level II	15*51/125=6.12	51*8/125=3.26
Level III	25*80/174=11.49	48*21/174=5.79
Level IV	67*35/166=14.13	14*50/166=4.22
Level V	43*10/98=4.39	8*37/98=3.02
Σ	**37.13**	**16.29**

$$\alpha_{MH} = \frac{\sum_{i=1}^{n} \frac{A_i D_i}{N_i}}{\sum_{i=1}^{n} \frac{B_i C_i}{N_i}} = \frac{37.13}{16.29} = 2.28$$

The item presents DIF. It should be removed to avoid the discrimination observed against Andalusian students.

88

7. Summary

- Psychometric characteristics that a good item should present (apart from relevance and representativeness, e.g.):
  - Difficulty (in tests to measure ability):
    - Between 0.2 and 0.8.
    - Most of them should be between 0.3 and 0.7.
  - Discrimination:
    - In aptitude tests, at least over 0.3.
    - In attitude tests, at least over 0.2.

89

7. Summary

  - Distractors:
    - Have to be chosen for more participants with low scores in the test.
    - Have to be equally probable.
  - When participants with the same level in the construct to be measured present different probability of answering correctly an item, a procedure to detect DIF have to be carried out. If the item presents DIF, it should be reviewed or removed.

90