Evaluating the effectiveness of multiple-choice assessment using item response curves

Bojan Golli

Pedagoška fakulteta, UL

IPRIF, April 8, 2021

Motivation

Multiple choice tests

Advantages

- easy to create, score and analyse.
- more affordable for testing a large number of students.

Disadvantages

- limited types of knowledge that can be assessed
- partial understanding of the subject
- selecting a random answer

Example

Consider the following formulation:

When a rubber ball dropped from the rest bounces off the floor, its direction is reversed because,

1. energy of the ball is conserved
2. momentum of the ball is conserved
3. angular momentum of the ball is conserved

33% without any knowledge
ev. 50% by eliminating the obviously wrong answer

So how to design and asses a good multiple choice question (MCQ)?

Comment: national competition for the first year of the secondary school (Čmrlj):
correct answer: 4 points
four wrong answers: -1 point
no answer: 0 point

Outline

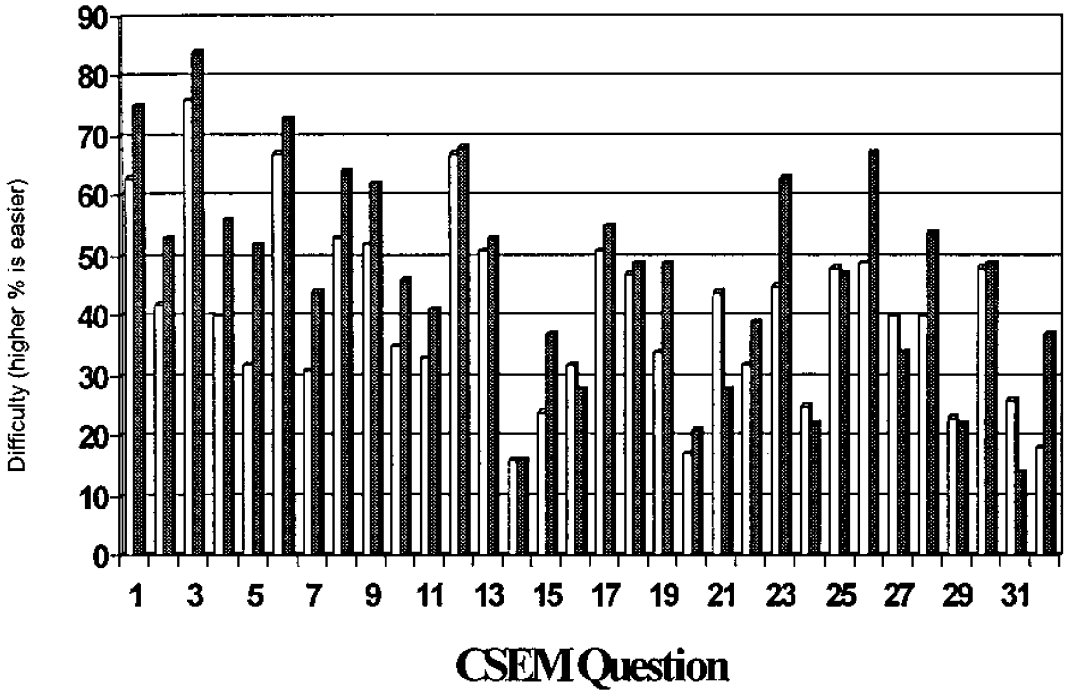
- How to asses (Criteria for assesing) multiple choice questions: *difficulty, discrimination, validity, effectiveness, reliability*
- Item response curves: a simple technique for evaluating MCQ
- Examples from *Force Concept Inventory* (FCI)
- Examples from national competitions for elementary school
- How to extract relevant data from the results of the national competition Čmrlj (Bumblebee, Bumbar) and how to prepare the IR curves.

Criteria for assesing multiple choice question

- difficulty (*težavnost*)
- discrimination (*ločljivost*)
- validity (*veljavnost*)
- reliability (*zanesljivost*)
- effectiveness (*učinkovitost*)

Difficulty

The percentage of subjects who get the item correct.
Ideal (average) 50 %, but reasonable range from 20 % to 80 %.
Example Survey of *conceptual knowledge of electricity and magnetism*



Discrimination

A measure of how well an item differentiates between competent and less competent students.

It is typically calculated as

$$\text{IDis} = \frac{N_U - N_L}{N_U + N_L}$$

where:
N_L the number of students in the bottom 27 % of the overall score
N_U the number of students in the top 27 % of the overall score

Discrimination values range from -1.0 to 1.0.

IDis $= 0.20$	traditional lower limit for acceptability
0 – 0,24	to be improved
0,25 – 0,39	good question
0,40 – 1,00	excellent question

Validity

An estimate of how well the test measures what it contends to measure.

Physics teachers/professors rate each item for both

- *reasonableness*:
  of question and items with respect to the curriculum
- *appropriateness*:
  with respect to students background in physics and mathematics

Reliability

The reliability is a measure of how consistently the test will reproduce the same score under the same conditions. The standard way to calculate the reliability of a test is to use Kuder-Richardson formula 20 (KR 20):

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_i\, q_i}{\sigma_X^2}\right] \qquad \sigma_X^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}$$

p_i ... is the proportion of correct responses to test item i,
q_i ... is the proportion of incorrect responses to test item i ($p_i + q_i = 1$);
K ... is number of questions and
N ... number of students.

Values in the range 0.8 to 0.9 are very high and indicate a test that can be used for both individual and group evaluation. Values in the range 0.7 to 0.8 are common for well-made cognitive tests. Values in the range 0.6 to 0.7 are considered weak for cognitive tests, but are acceptable for personality tests. A range of 0.5 to 0.6 is common for well-made classroom tests.

Effectiveness (Učinkovitost)

How effective are the distractors (incorrect answers)?

Distractors are usually based on common misconceptions (in mechanics, from *Force Concept Inventory*):

- impetus, active force
- action/reaction pairs
- concatenation of influences: one force wining over the other
- other influence on motion

Item response curves

IRC a simplified version of item response theory

probes student understanding as a function of ability level through an examination of each answer choice

The procedure is illustrated on a problem from national competition for elementary school

Item response curves

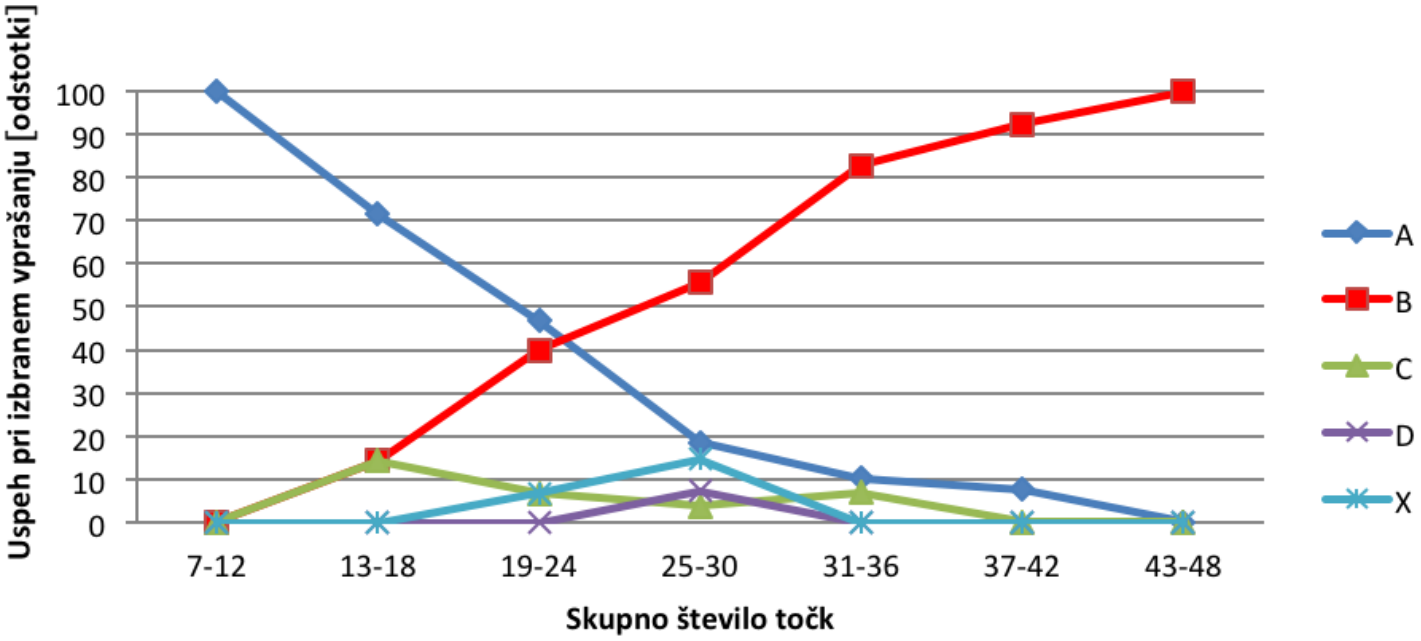
Ladja zapluje iz slanega morja v reko. Kaj se zgodi z vzgonom in ladjo?

- A. Vzgon se poveča, ker se ladja bolj pogrezne.
- B. Vzgon se ne spremeni, čeprav se ladja pogrezne.
- C. Vzgon se zmanjša, ker se ladja nekoliko dvigne.
- D. Vzgon se ne spremeni, čeprav se ladja nekoliko dvigne.

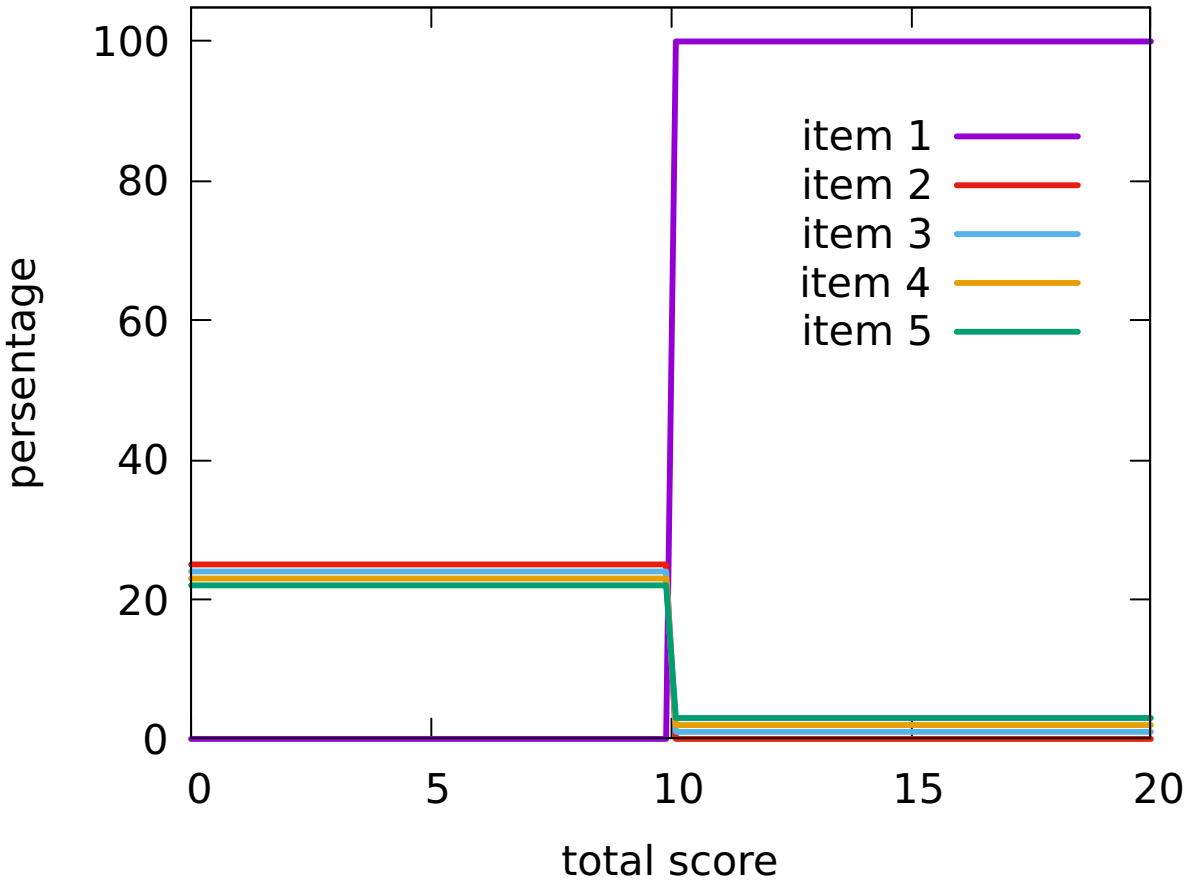
score	Nstud	A	**B**	C	D	X	A	**B**	C	D	X
7–12	1	1	**0**	0	0	0	100	**0**	0	0	0
13–18	7	5	**1**	1	0	0	71.4	**14.3**	14.3	0	0
19–24	15	7	**6**	1	0	1	46.7	**40**	6.7	0	6.7
25–30	27	5	**15**	1	2	4	18.5	**55.6**	3.7	7.4	14.8
31–36	29	3	**24**	2	0	0	10.3	**82.8**	6.9	0	0
37–42	13	1	**12**	0	0	0	7.7	**92.3**	0	0	0
43–48	4	0	**4**	0	0	0	0	**100**	0	0	0

Item response curves

score	Ns	A	**B**	C	D	X	A	**B**	C	D	X
7–12	1	1	**0**	0	0	0	100	**0**	0	0	0
13–18	7	5	**1**	1	0	0	71.4	**14.3**	14.3	0	0
19–24	15	7	**6**	1	0	1	46.7	**40**	6.7	0	6.7
25–30	27	5	**15**	1	2	4	18.5	**55.6**	3.7	7.4	14.8
31–36	29	3	**24**	2	0	0	10.3	**82.8**	6.9	0	0
37–42	13	1	**12**	0	0	0	7.7	**92.3**	0	0	0
43–48	4	0	**4**	0	0	0	0	**100**	0	0	0



Ideal case



All distractors have (almost) equal probability (25 %)
Sharp transition between students ability levels.

Features

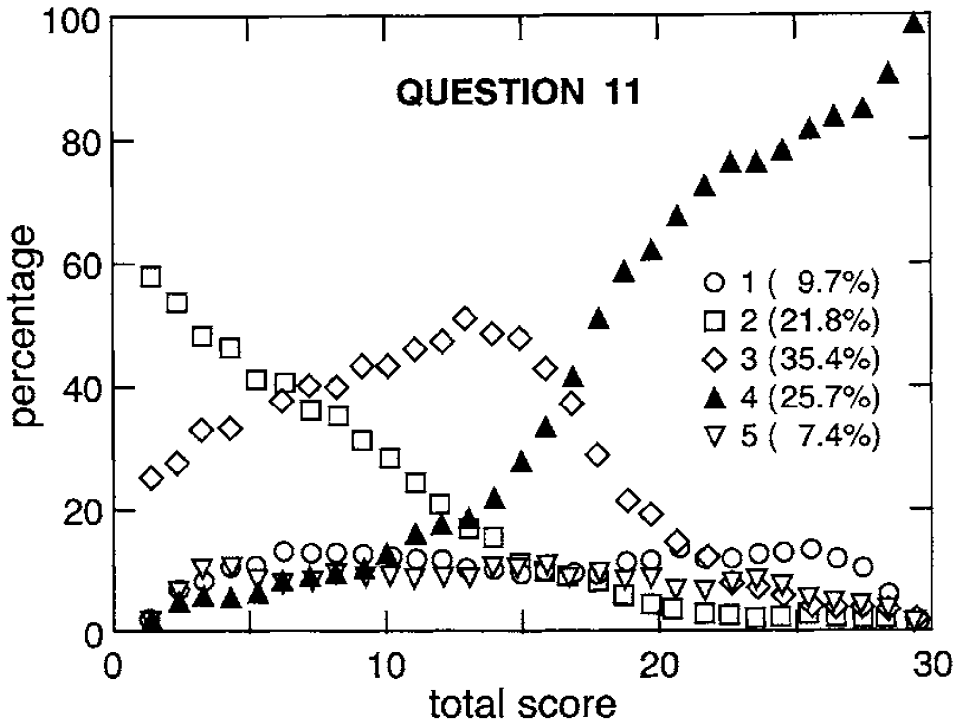
- identifying nonfunctioning distractors that can be replaced with distractors attractive to students at various ability levels.
- identify prominent misconceptions
- tailor instructions to combat those misconceptions,
- possibility to discriminate medium and low level ability:
- the sharp slope identifies an item as highly discriminating.

Illustrated examples from the item response analysis of *Force Concept Inventory*

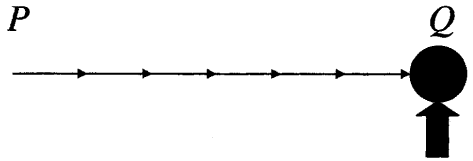
Example of a difficult but efficient question

The main force(s)acting on the puck after receiving a kick is (are):

1. a downward force of gravity.
2. a downward force of gravity, and a horizontal force in the direction of motion.
3. a downward force of gravity, an upward force exerted by the surface, and a horizontal force in the direction of motion.
4. a downward force of gravity and an upward force exerted by the surface.
5. none.

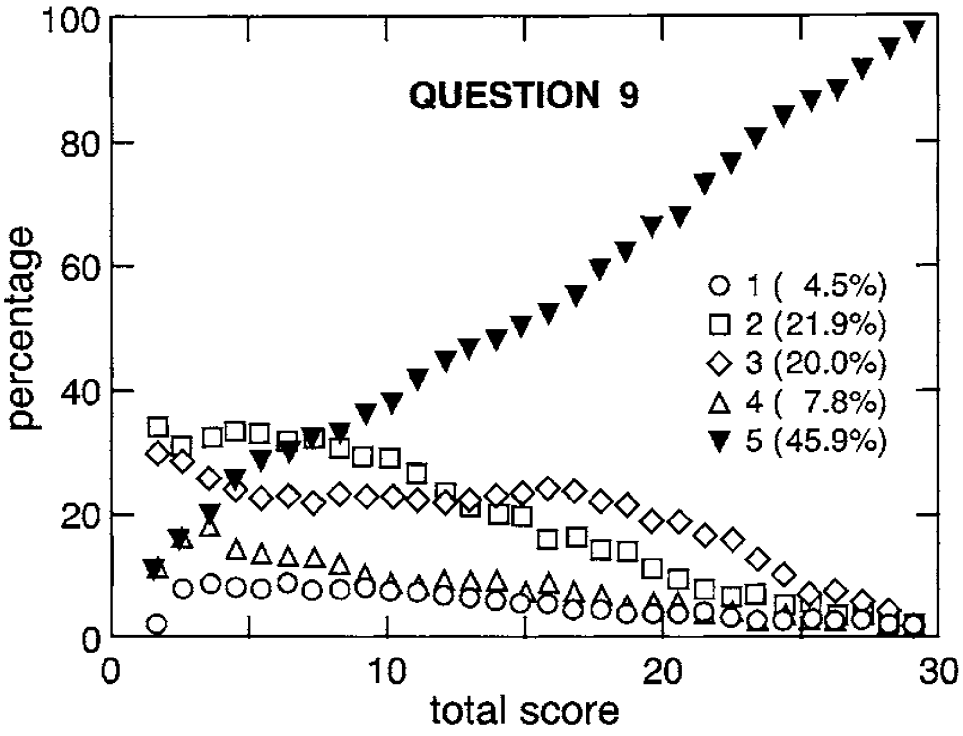


Moderately efficient



The speed of the puck just after it receives the kick is

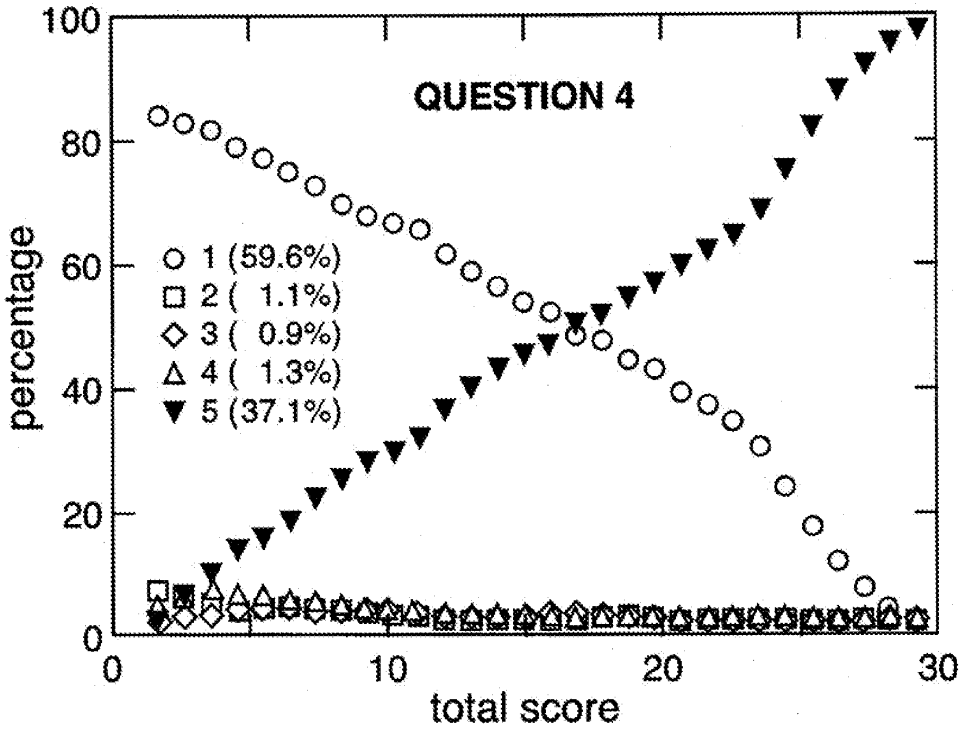
1. equal to the speed v_0 it had before it received the kick.
2. equal to the speed v_k resulting from the kick and independent of the speed v_0 .
3. equal to the arithmetic sum of the speeds v_0 and v_k .
4. smaller than either of the speeds v_0 or v_k
5. greater than either of the speeds v_0 or v_k , but less than the arithmetic sum of these two speeds.



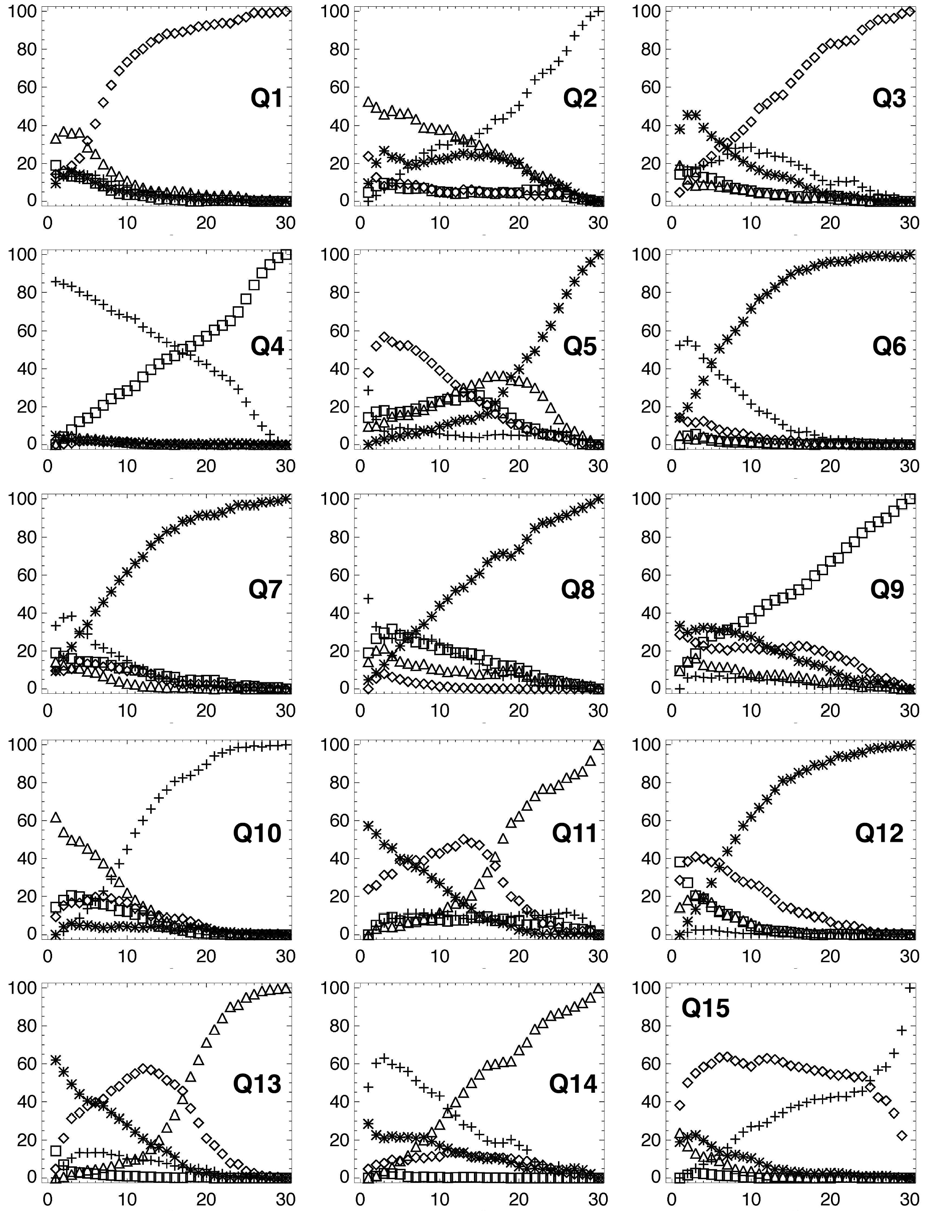
Moderately difficult but nonefficient

A large truck collides head-on with a small compact car. During the collision

1. the truck exerts a greater amount of force on the car than the car exerts on the truck.
2. the car exerts a greater amount of force on the truck than the truck exerts on the car.
3. neither exerts a force on the other, the car gets smashed simply because it gets in the way of the truck.
4. the truck exerts a force on the car but the car does not exert a force on the truck.
5. the truck exerts the same amount of force on the car as the car exerts on the truck.



Analysis of FCI Q1-15



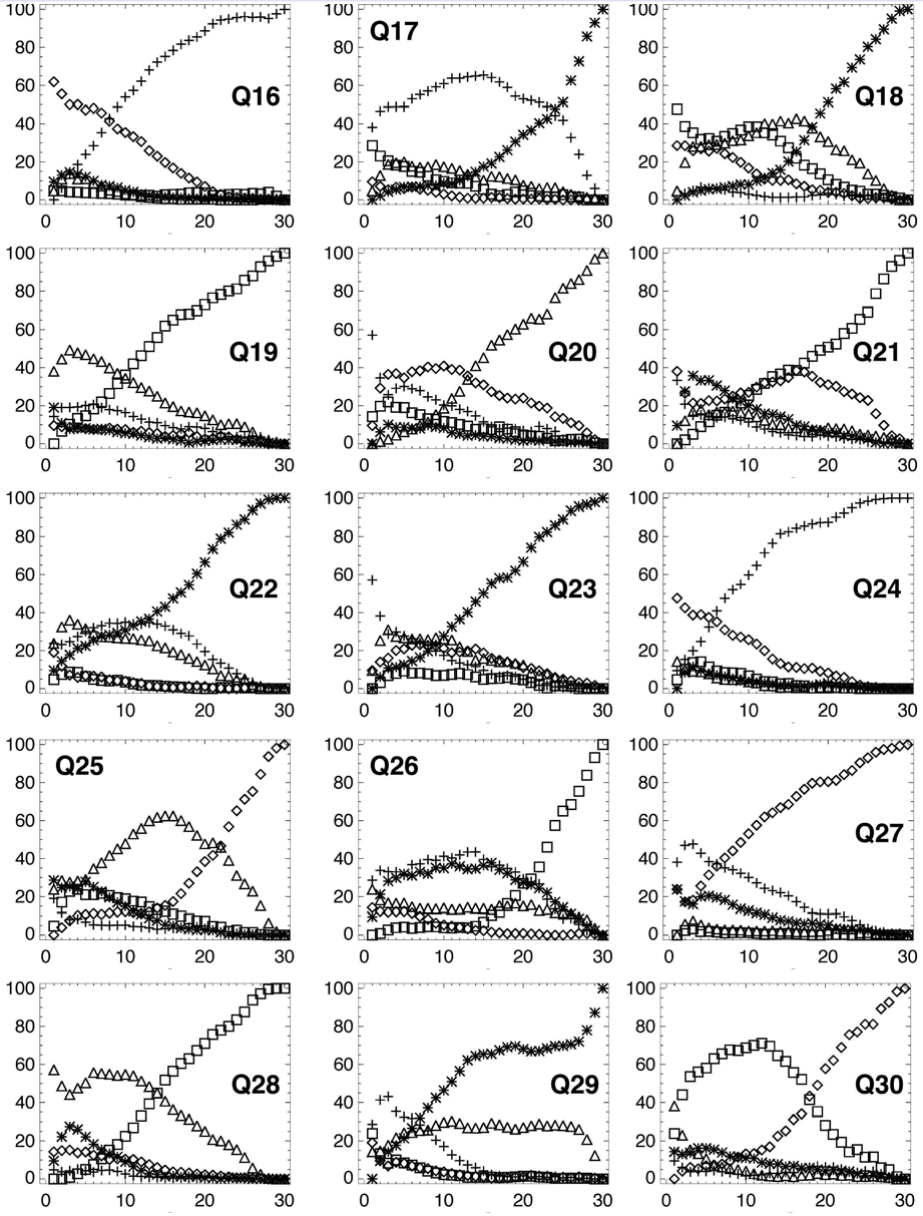
Q1: Improvement could be made by changing one of the low-appeal distractor choices to be more attractive to moderate ability students.

Q5: All distractors functioning reasonably well.

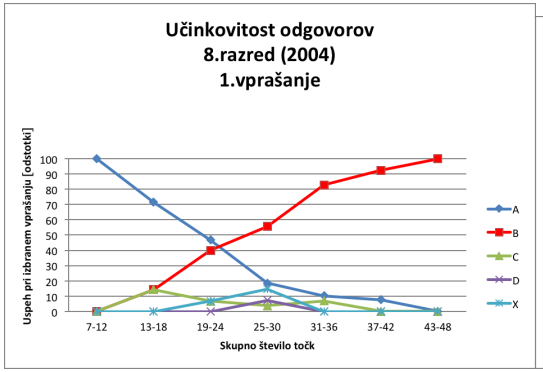
Q6: Two distractors inefficient.

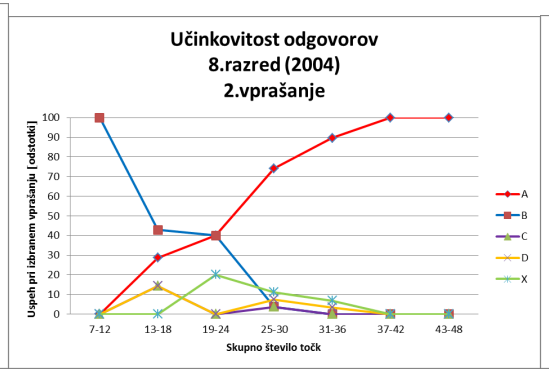
Q13: Could be used to characterize the ability level of a student into one of three ranges.

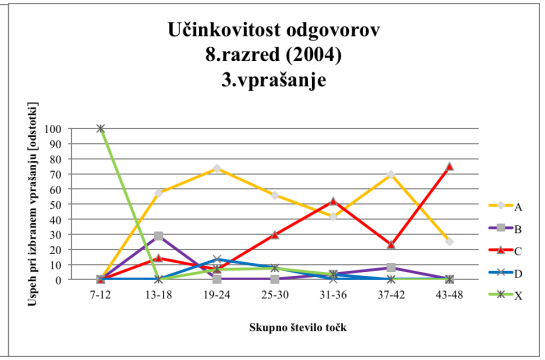
Analysis of FCI Q16-30

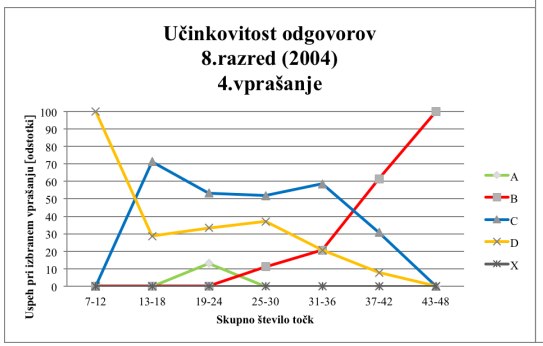


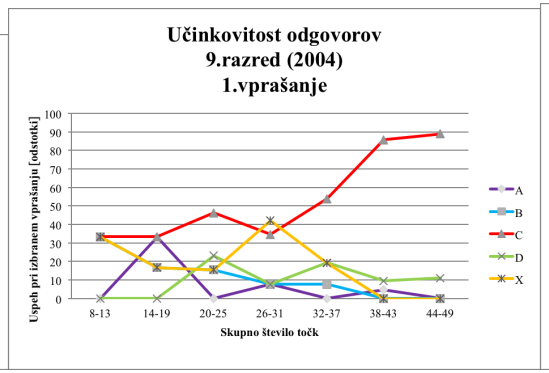
National competition for elementary school

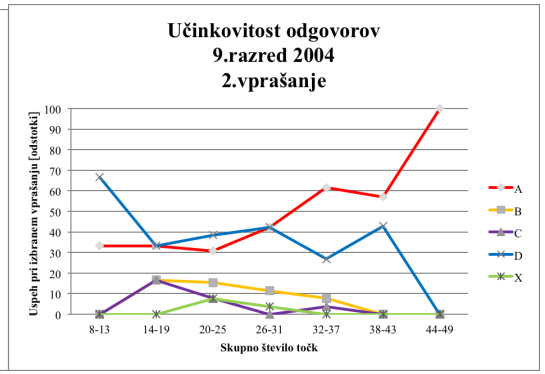


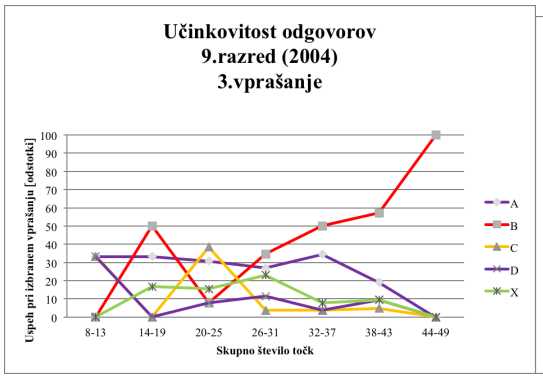


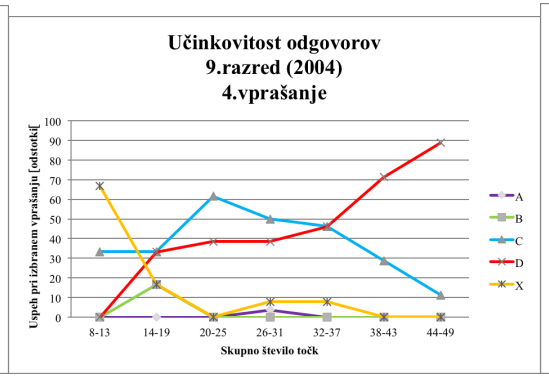


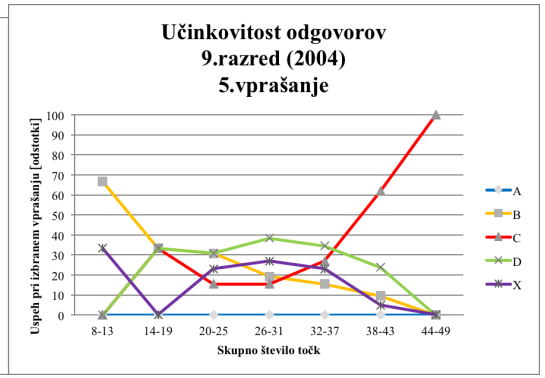












Literature

- D. Hestnes *et al.*, *Force Concept Inventory* The Physics Teacher, Vol 30, March 1992, 141.
- D. P. Maloney *et al.*, *Surveying students' conceptual knowledge of electricity and magnetism,* Phys. Educ. Res., Am. J. Phys. Suppl., Vol. 69, No. 7, July 2001, S12.
- G. A. Morris *et al.*, *Testing the test: Item response curves and test quality,* Am. J. Phys., Vol. 74, No. 5, May 2006, 449.
- Gary A. Morris et. al *An item response curves analysis of the Force Concept Inventory* Am. J. Phys. 80 (9), September 2012
- J. D. Marx *et al.*, Am. J. Phys., Vol. 75, No. 1, January 2007, 87.

Seminar

From a particular field of (primary school) physics choose five or more questions and perform the IRC analysis

Comment on:

- rank the questions with respect to their efficiency (discrimination)
- explain criteria used for ranking
- identify nonfunctioning distractors
- identify well-functioning and attractive distractor
- identify most common misconceptions