

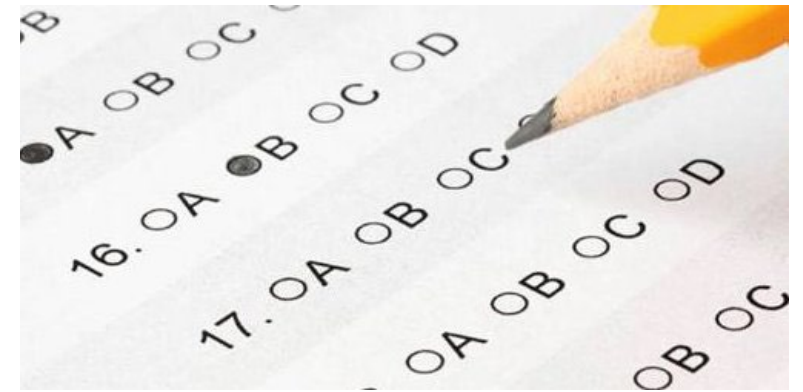
Elimination Marking

A new standard for multiple choice exams?

Disclaimer

The bad reputation of MC tests is unwarranted*:

- efficient
- objective
- reliable
- content validity
- exam quality safeguarding



* when comprised of well-designed questions

Ezra

A woman visits the doctor

What is wrong with her?

- Diagnosis A
- Diagnosis B
- Diagnosis C
- Diagnosis D

What would you have the doctor do?



Erasmus

A woman visits the doctor

What is wrong with her?

- Diagnosis A
- Diagnosis B
- Diagnosis C
- Diagnosis D



standard setting

- ~~Diagnosis A~~
- Diagnosis B
- ~~Diagnosis C~~
- Diagnosis D

elimination marking

Standard setting

- MC questions with 4 alternatives
- Students answer all questions
- Guessing chance of 25%

➤ raise standard for passing



choice	A	B	C	D	score
the one correct alternative	Ⓐ	●	ⓒ	Ⓓ	1
one of the incorrect alternatives	Ⓐ	Ⓑ	●	Ⓓ	0

Negative marking (guess correction)

- MC questions with 4 alternatives
- Students choose which questions to answer
- Penalty for mistakes



choice	A	B	C	D	X	score
the one correct alternative	Ⓐ	●	ⓒ	Ⓓ	⊗	1
one of the incorrect alternatives	Ⓐ	Ⓑ	●	Ⓓ	⊗	-0,333
blank	Ⓐ	Ⓑ	ⓒ	Ⓓ	●	0

Risk of extraneous influences

Level playing field?



- Risk aversion: female > male
 - When in doubt, female students leave questions unanswered
- Test anxiety: female > male
 - Anxious individuals do worse on MC ~ information processing

Ezra

Risk of extraneous influences

Level playing field?



'Afschaffen giscorrectie maakt examens niet eerlijker'

GISCORRECTIE, DE FOUTE KEUZE?

Giscorrectie benadeelt 'eerlijke' studenten

'Examens worden niet eerlijker door giscorrectie af te schaffen'

Standard setting: no extraneous influences

Erasmus

Psychometric critique

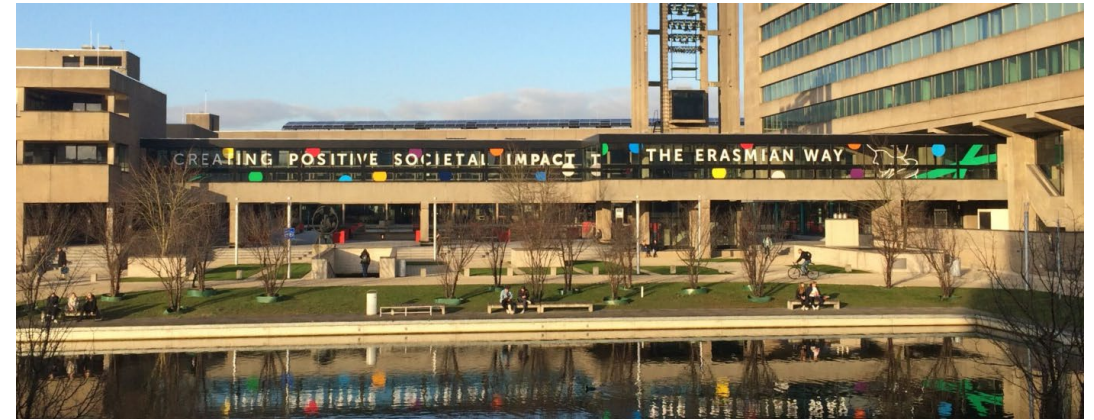
High-stakes tests

- students guess to different degrees
 - no distinction between blind and informed guessing
 - guessing introduces extraneous noise
-
- attenuates test reliability
 - complicates the interpretation of observed scores as true measures of competence



Ezra

Educational critique



Misaligned incentives

- risk: strategic expected-value maximizers
- signal: luck is a valid substitute for competence
- goal: expertise and professional accountability
- goal: accurately diagnose the limits of one's own knowledge

Limited information about learning process

- students: miscalibration and overconfidence
- teachers: how to support individual learning?

Erasmus

Elimination marking

- Students eliminate distractors they believe to be incorrect
- Students are rewarded proportionally to their ability to rule out faulty alternatives

1.0p 4 Brown and Lenneberg (1954) presented students with a series of 24 color chips. The students were asked to name the colors one at a time. Brown and Lenneberg measured the time it took the students to come up with a name. Brown and Lenneberg found that colors that evoked long names were named with hesitation. That is, less codable colors elicited longer reaction times.

This finding supports

- ☐ the strong version of linguistic determinism.
- ☐ the weak version of linguistic determinism.
- ☐ ~~the strong version of linguistic relativity.~~
- ☐ ~~the weak version of linguistic relativity.~~

Elimination marking

- Students eliminate distractors they believe to be incorrect
- Students are rewarded proportionally to their ability to rule out faulty alternatives

[A* B C D]	knowledge type	score	response behavior
[0 1 1 1]	perfect knowledge	1	3 distractors eliminated
[0 0 1 1]	partial knowledge	1/3	2 distractors eliminated
[0 0 0 1]	partial knowledge	1/9	1 distractor eliminated
[0 0 0 0]	no knowledge	0	no distractors eliminated
[1 * * *]	misconception	-1/3	correct answer eliminated

A stylized, handwritten-style logo of the word "Erasmus" in a dark blue or black color.

Benefits

- Practical: natural test taking process
- Validity: partial knowledge is recognized
- Reliability: fewer noise → more accurate grade
- Educational: more insight in learning process
- Educational: signals professional behavior

Ezra

Worries

More complex and less transparent

- Cognitively taxing and stress inducing

Reduces but does not eliminate guessing/strategic responding

- Role of extraneous influences ~ gender, culture
- Grade inflation/deflation

➤ Do psychometric and pedagogical benefits outweigh potential practical disadvantages?

Procedure

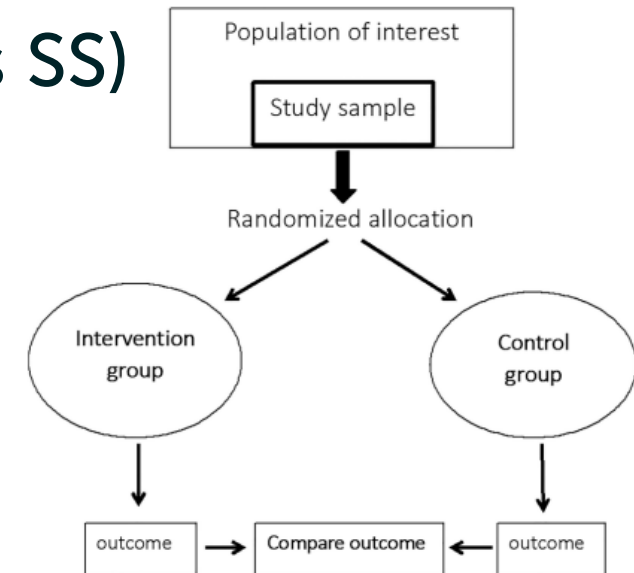
- Practice exam Cognitive Psychology
- First knowledge course of BA2
- Halfway the course/block
- 20 former exam questions in English
- MC with 4 answer alternatives, 1 correct



Ezra

Procedure

- Students randomly assigned to SS or EM
- Instructed beforehand and at start of practice exam
- Test anxiety, risk-taking propensity, gender, nationality
- Test taking experience (7 questions, 5-point Likert scale)
- BA1 grade + actual exam result (40 MC questions SS)
- Interviews with EM students



Procedure

Two EM approaches:

- Exclusion scoring: eliminate all answers known to be incorrect
- Inclusion scoring: create smallest subset of answers of which the student is confident contains the correct answer

Opposite framing, similar grading



Sample

- 299 students (49%) → 276 after exclusions (45%)
- Average age 21 years
- 82% women, 13% men, 5% other
- 94 Dutch, 182 international track (includes Dutch students)

Sample

- 299 participants (49%) > 276 after exclusions (45%)
- Average age 21 years
- 94 Dutch, 182 international track (includes Dutch students)
- 82% women, 13% men, 5% other
- Randomly assigned to SS (134) or EM (142)
- No difference between SS and EM in terms of gender, nationality, test anxiety, risk taking, BA1, course exam

Ezra

Sample

- 299 participants (49%) > 276 after exclusions (45%)
- Average age 21 years
- 94 Dutch, 182 international track (includes Dutch students)
- 82% women, 13% men, 5% other
- Women ($M = 2.34$) more test anxiety than men ($M = 1.84$)
- Women ($M = 2.76$) less risk taking than men ($M = 3.05$)
- No effect of cohort (Dutch/International)

Ezra

Results: BA1

- No gender effect
- Cohort effect (internationals better)
- Correlation with test anxiety: $-.199^*$
- Correlation with risk taking propensity: $-.124^*$



Results: Multiple option use EM

- Students in the EM condition selected significantly more answer options ($M = 1.33$) than students in the SS condition ($M = 0.99$)
- 93% of students used EM option and did so on 25% of questions (median)
- Considerable individual differences (5% - 85% of questions)
- No relationship with BA1 grade, test anxiety, risk taking, gender or cohort
- Less options for easy questions ($r = -.66^*$)

A stylized, handwritten-style logo of the word "Erasmus" in a dark blue or black color.

Results: Grades and passing rates

- No significant difference in mean grades (EM = 4.62 vs SS = 4.99)
- No significant difference in passing rate (EM = 37% vs SS = 43%)
- SS predicts exam grade better (EM = .164 vs SS = .472)
- SS correlates better with BA1 grade (EM = .202 vs SS = .358)
- For comparison: exam grade and BA1 grade correlation = .552

A handwritten signature in black ink, appearing to read 'Erasmus', is located in the bottom right corner of the slide.

Results: Demographics

- No effect of gender in EM or SS
- Effect of cohort in both EM and SS (international better)

	Test anxiety	Risk taking
BA1 grade	-.199	-.124
Exam grade	-.207	-.204
EM	-.206	-.239
SS	-.315	-.086 [†]

Results: Examination experience

	EM	SS
Confusing	2.42	2.48
Fair	3.11	3.29
Critical thinking	3.08	3.24
Confidence answers	2.46	2.45
Stress	2.51	2.40
Representative grade	2.67	2.85
Confidence exam	2.65	2.77

Results: Qualitative feedback

Advantages:

- ability to express uncertainty, especially when in doubt between 2 alternatives ("safer")
- adequate representation of knowledge

Disadvantages:

- difficulty adjusting to new format
- increased stress due to inexperience with format
- Increased stress due to unclear impact of strategic responding

A stylized, handwritten-style logo of the word "Erasmus" in a dark blue or black color, located in the bottom right corner of the slide.

Conclusions EM

The good

- Explicit display of partial knowledge
- No grade inflation/deflation
- Gender/cohort neutral
- No more pronounced relationship with external variables
- Subjective experience on par with SS
- Students highly appreciated practice opportunity

The Erasmus logo, featuring the word "Erasmus" in a stylized, cursive script.

Conclusions EM

The bad

- More complex format that requires getting used to
- More strategic trends with experience?
- Ans not currently set up for EM
- Grading complexity
- Test anxiety has a detrimental effect overall

Erasmus

Conclusions EM

The ugly

- Practice test on voluntary basis for which many students were ill prepared
- Specific content (abstract, somewhat technical)
- Limited number of male students
- Overrepresentation international students
- Extensive experience with “default” SS
- Inclusion scoring rather than exclusion scoring
- Test anxiety and risk taking measured after test

Contributors

Who contributed the most?

- A. Junior Louter
- B. Julia Perik
- C. Samuel van Zuijlen
- D. Oliver Lindemann
- ~~E. Steven Verheyen~~



**NATIONAAL REGIEORGAAN
ONDERWIJSONDERZOEK**

What is it we want?

SS

- Simpler for students and staff
- Easier to administer

EM

- More reliable
- More didactically appropriate

