Questionmark
powered by Learnosity

Item Analysis for Beginners

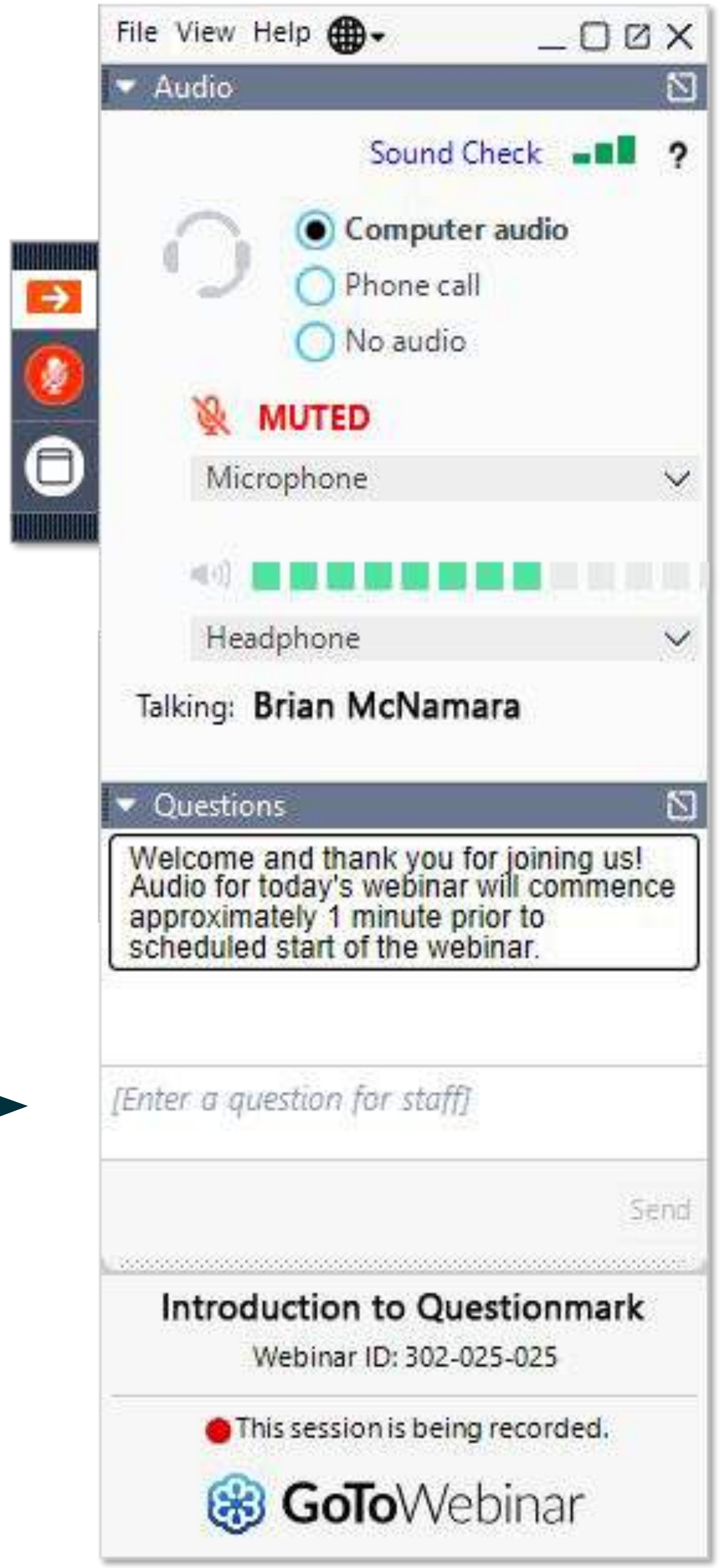
May 24, 2023

Before we get started

Watch for an email after the webinar to:

- Download slides (PDF)
- View a recording
- Explore valuable resources

To ask questions, use the “Questions” feature



● Computer audio
○ Phone call
○ No audio

MUTED

Microphone

Headphone

Talking: Brian McNamara

Welcome and thank you for joining us! Audio for today's webinar will commence approximately 1 minute prior to scheduled start of the webinar.

[Enter a question for staff]

Introduction to Questionmark

Webinar ID: 302-025-025

This session is being recorded.

GoToWebinar

John Kleeman

EVP and Founder
of Questionmark

- John wrote the first version of the Questionmark assessment software system and founded Questionmark in 1988.
- Was on the original team that created the IMS Global Learning Consortium QTI specification and has worked on standards initiatives with ADL, AICC, ATP, BSI, ISO and others.
- EVP of Industry Relations and Business Development at Learnosity
- 30 years of experience in the assessment industry
- 2021 ATP Chairperson & current ATP Director

Who is this webinar for?

All who want to trust assessment results!

- This session is for you if:
  - you don't know about item analysis
  - you know you should do item analysis but haven't got round to it
  - you occasionally use item analysis but would like to know more
  - you are scared of statistics

Agenda

During this session, we will cover:

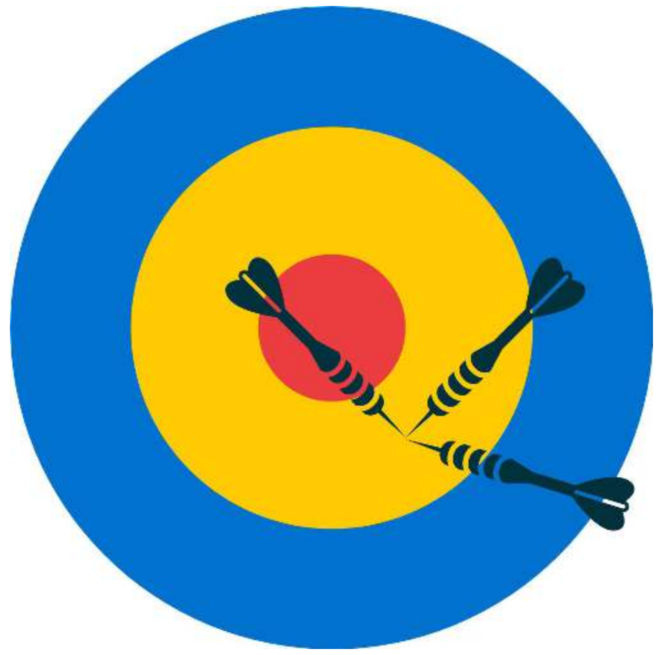
- Introductory concepts
- How easy or difficult a question is
- Looking at multiple choice distractors
- How well a question contributes to the assessment result
- Some practical exercises

Questionmark powered by Learnosity

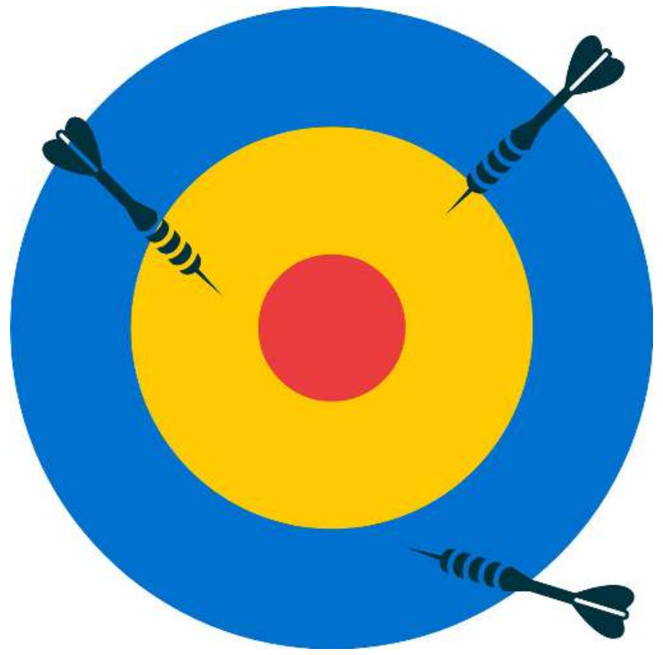
Reliable and Valid Assessments

- **Reliable:** dependable, repeatable, consistent
- **Valid:** measures appropriate knowledge and skills

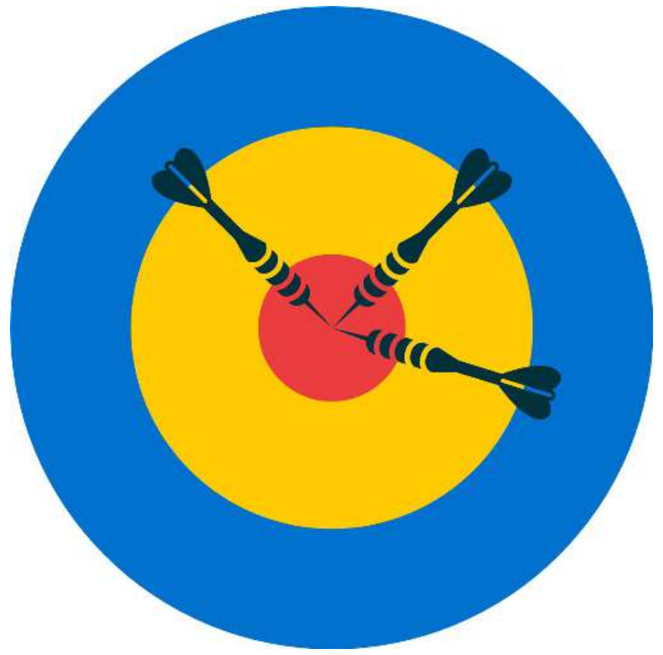
Item Analysis helps you get more reliable and valid



Reliable,
but not valid



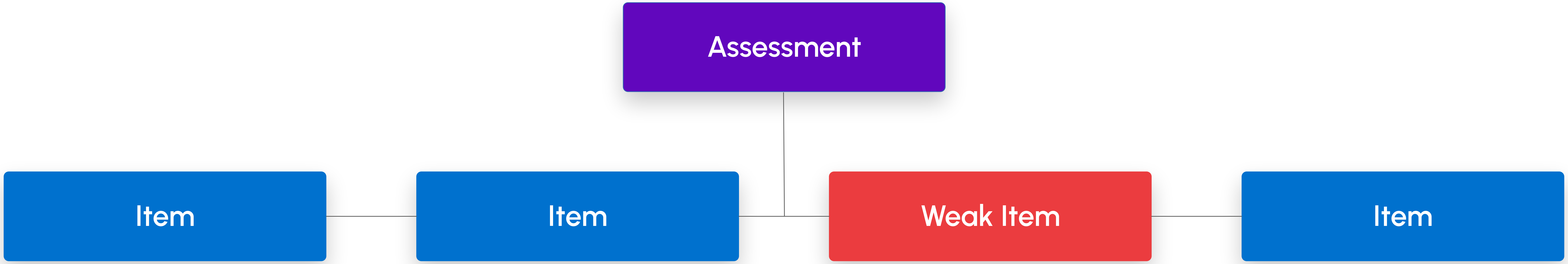
Not reliable and therefore
cannot be valid



Reliable
and valid

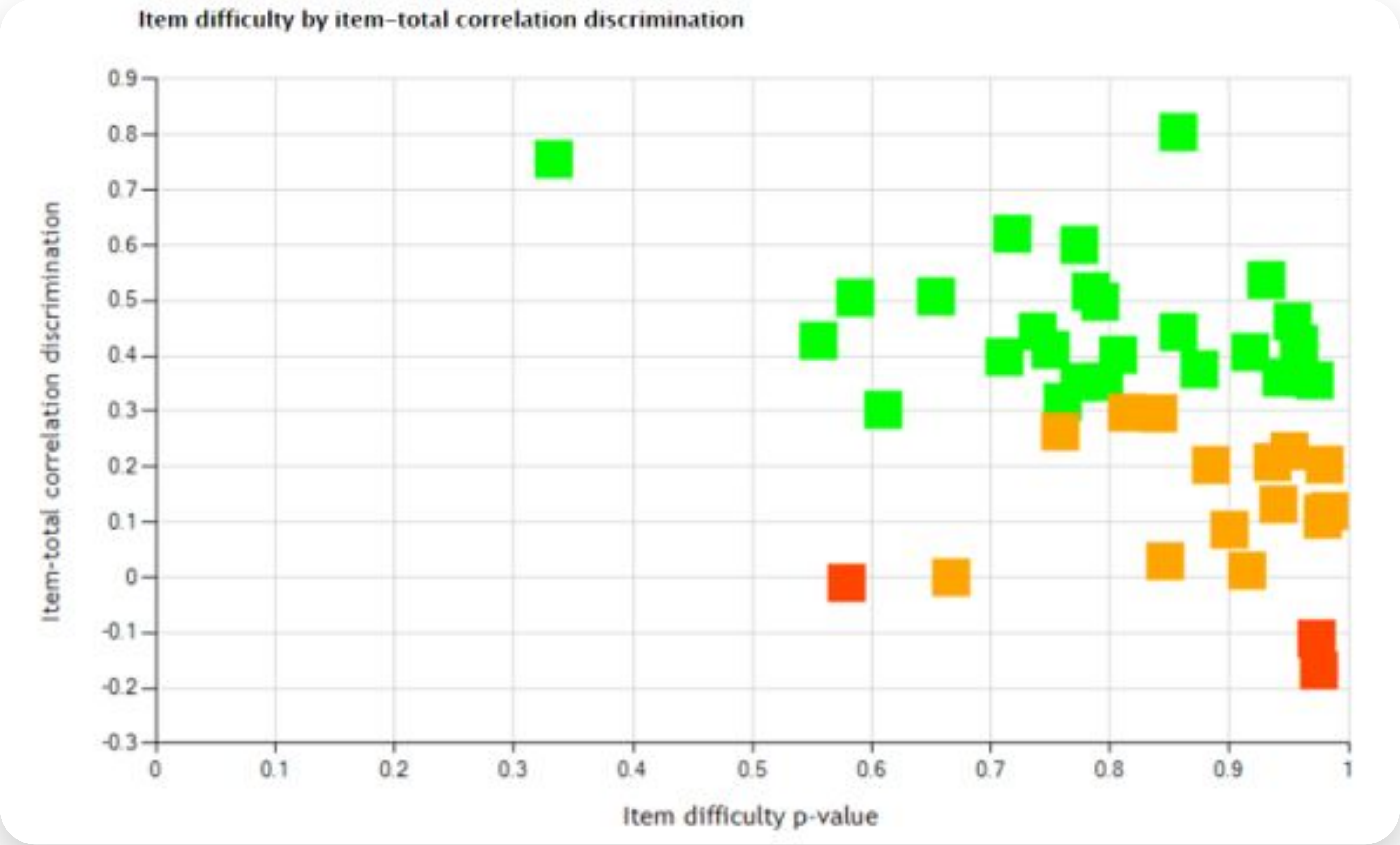
Questionmark
powered by Learnosity

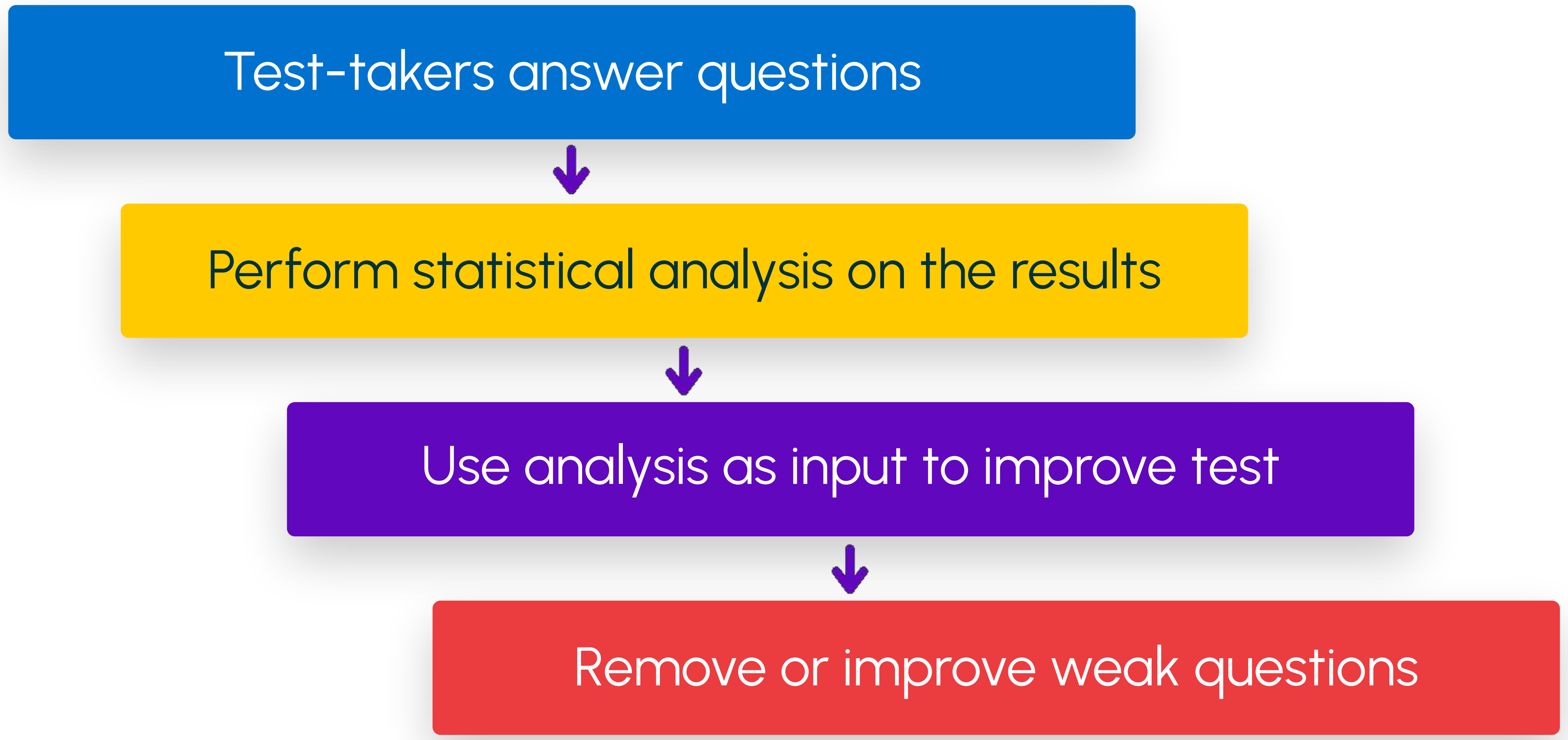
What is item analysis?



- An assessment contains many items (also called question)
- Item analysis looks at performance of each item
- To flag potentially weak items to remove or improve
- Removing/improving weak items makes assessment more reliable and valid

How it works





Test-takers answer questions

Perform statistical analysis on the results

Use analysis as input to improve test

Remove or improve weak questions

Questionmark
powered by Learnosity

How Item Analytics can help

Identify weak questions you can remove or improve

- Mis-keyed questions
- Ambiguous questions
- Irrelevant questions

Improve questions by removing weak distractors

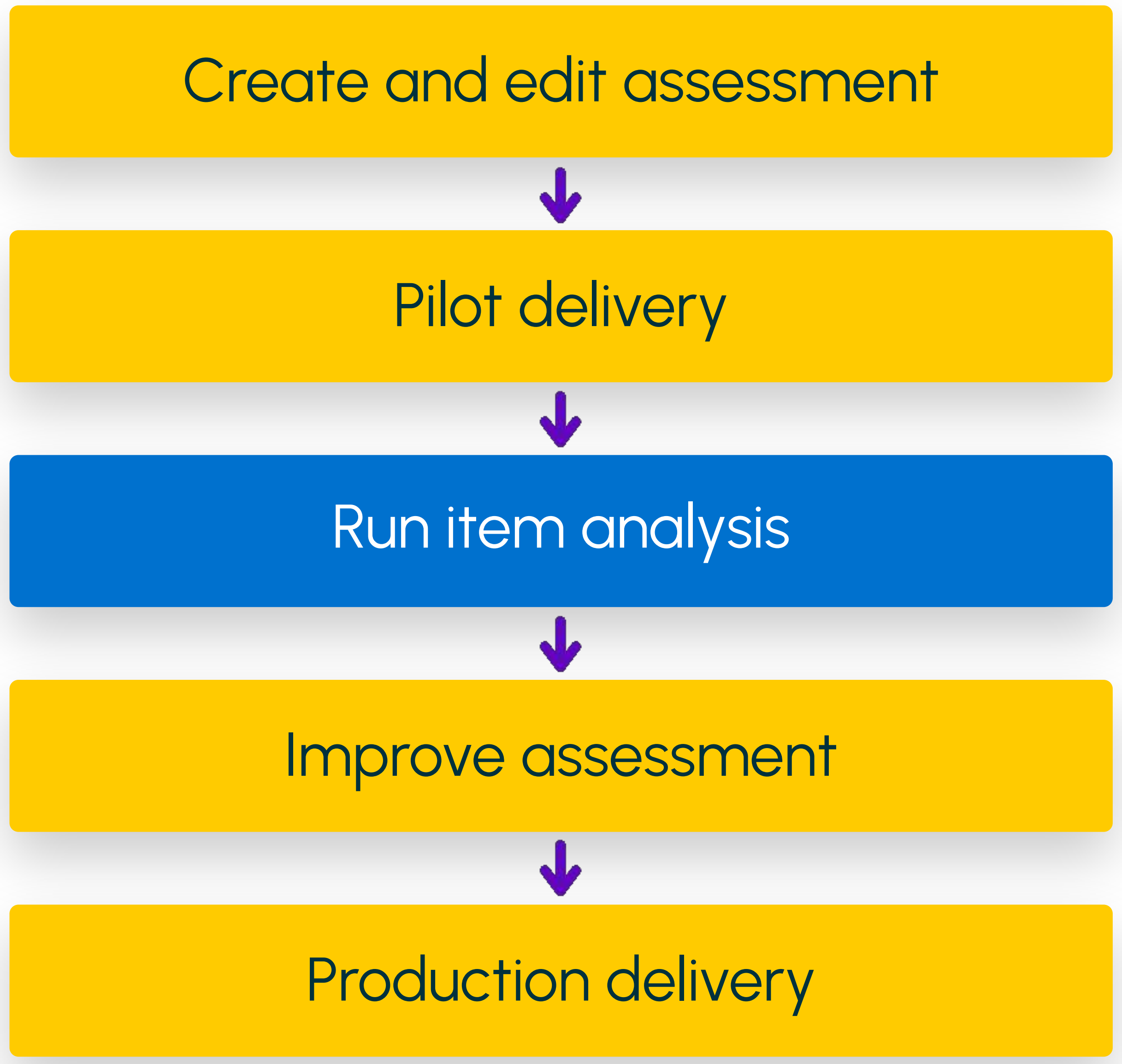
- Remove/change a choice no-one chooses
- Identify misleading or ambiguous choices
- Reduce ability to guess

Build confidence in your assessments

- Help make reliable, valid, fair and trustable
- Show stakeholders you follow good practice

When do you do Item Analysis?

After piloting a test



Create and edit assessment

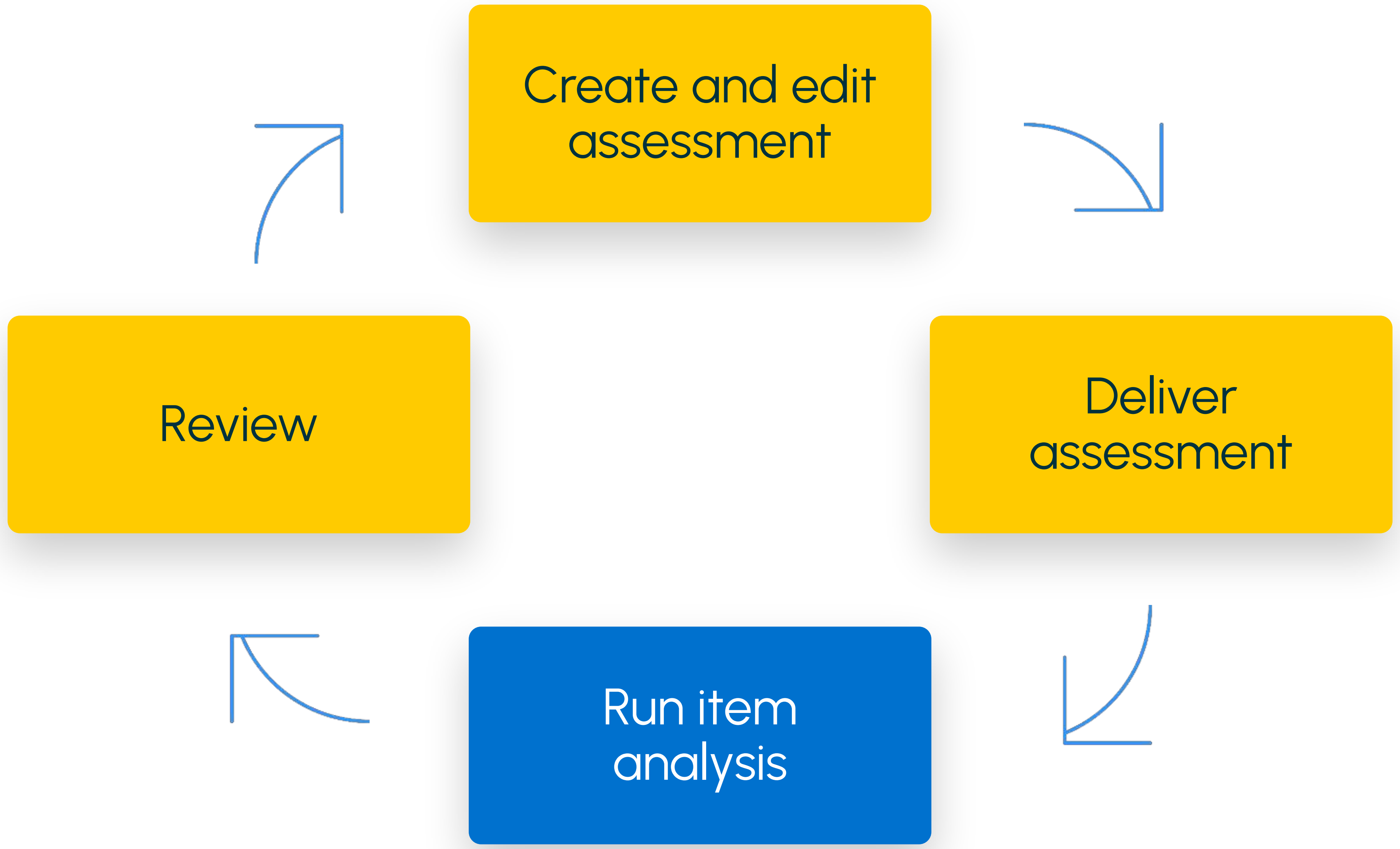
Pilot delivery

Run item analysis

Improve assessment

Production delivery

During monitoring of a production test



Create and edit assessment

Deliver assessment

Run item analysis

Review

Common questions

Is Item Analysis the same as Learning Analytics?

No. Item Analysis focuses particularly on the quality of questions for measurement.

Do I need an expert to do Item Analysis?

Although an expert can do more with it, it's useful for everyone.

How do I do Item Analysis?

- Built into many assessment systems including Questionmark.
- Can also do it by exporting data or in spreadsheets.
- Having 50+ results is helpful, 100+ best.

Is a question flagged by Item Analysis always bad?

- No. Item analysis not a magic wand, it highlights questions that might be weak or ambiguous.
- Bad items can have good stats and vice versa.
- Statistics also depend on the sample of results you are looking at.
- You also need to review items in other ways – e.g. content and bias.
- Item analysis helps identify items taking up unnecessary space in your assessments or that may weaken your assessment.

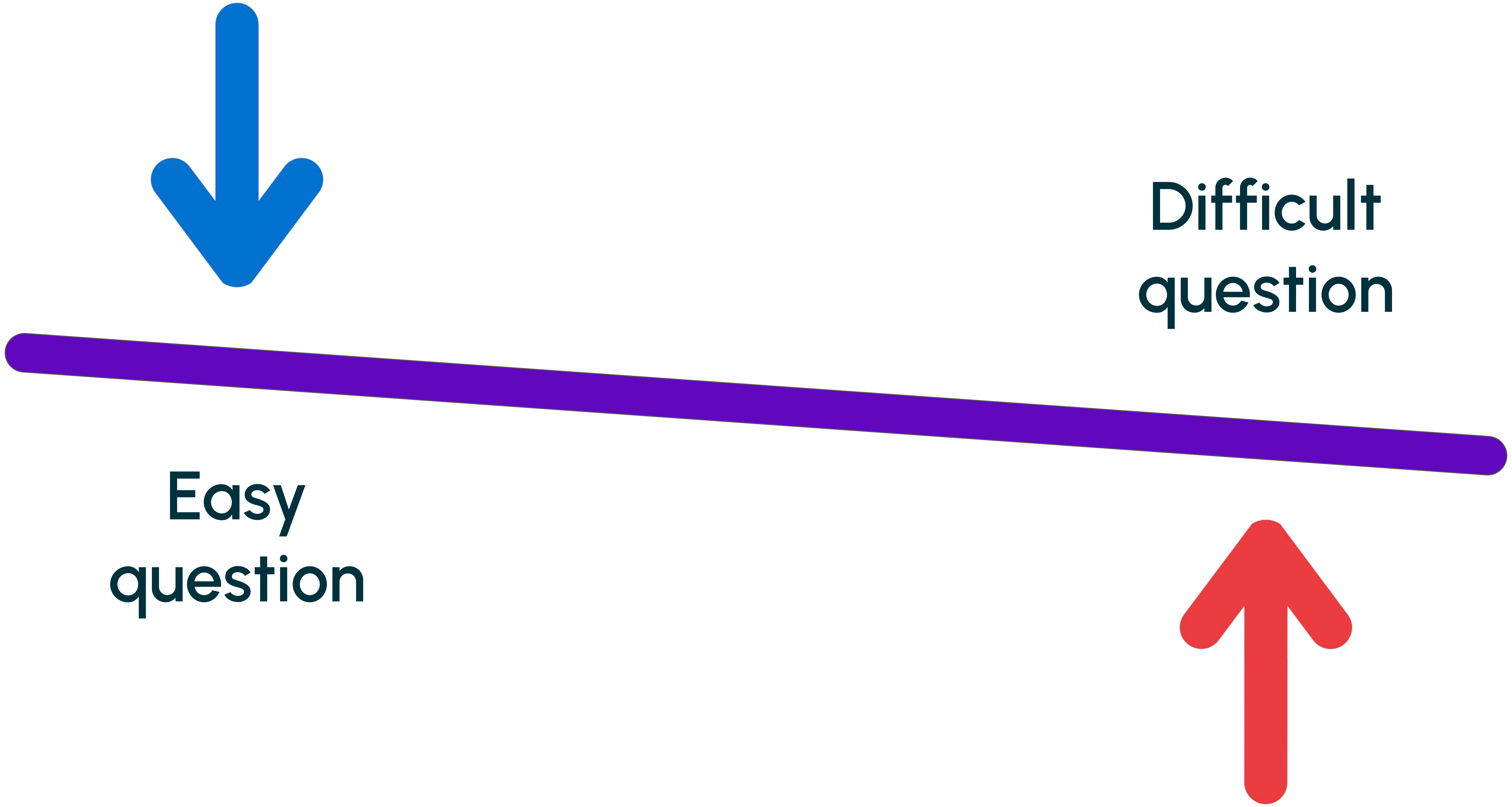
Does the webinar cover everything on item analysis?

No. It is "Item Analysis for Beginners" and there is more you can learn.

Question Difficulty:
How easy or difficult an item is

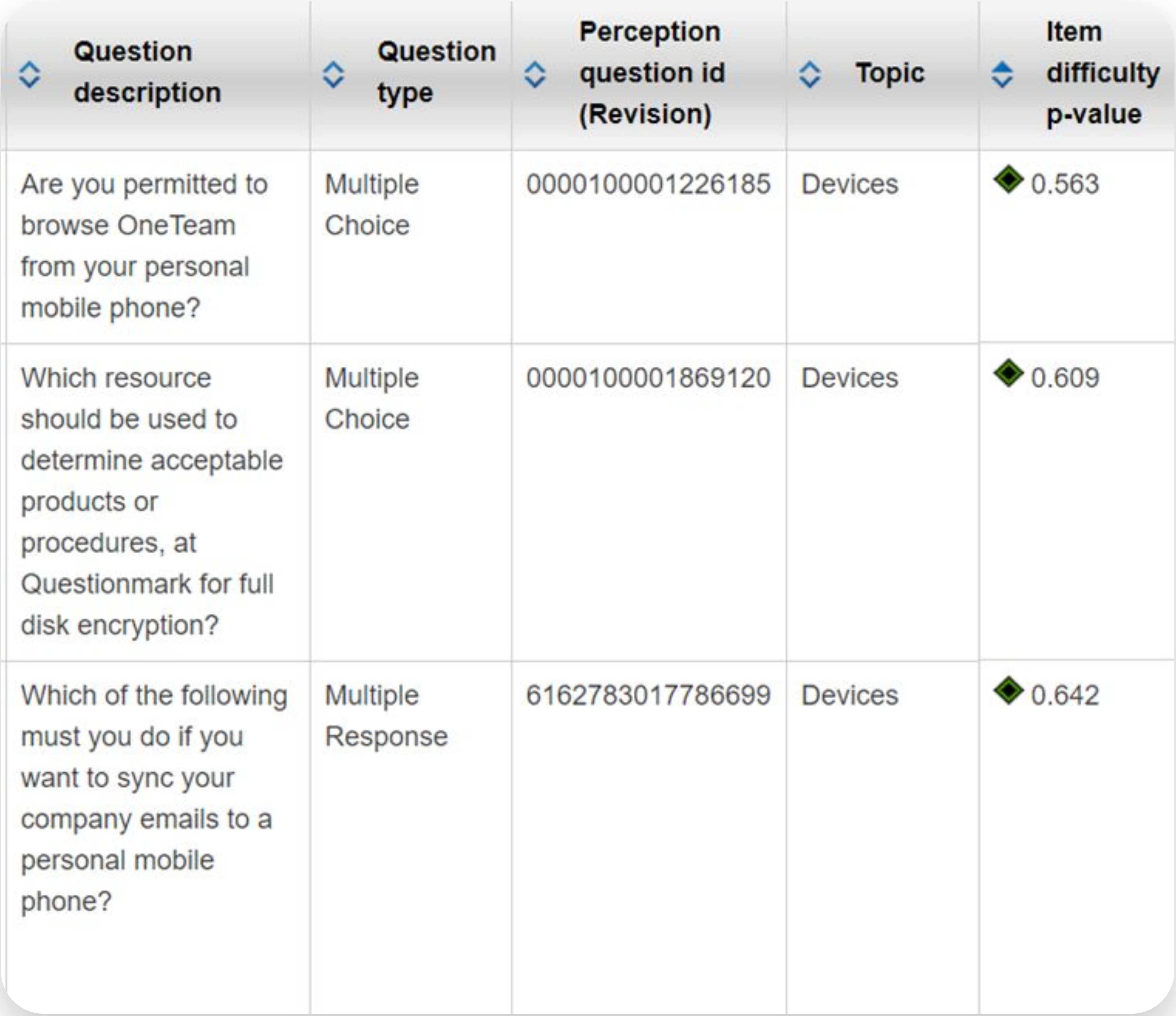
Technique 1 of 3

What do we mean by an Easy or Difficult question?



How We Measure Difficulty

- We use a sample of the people who have taken the assessment
- The item analysis report calculates "**p-value**" as a number between 0 and 1
- Roughly % of people who get a question right
- The higher the p-value, the easier a question is



Question description	Question type	Perception question id (Revision)	Topic	Item difficulty p-value
Are you permitted to browse OneTeam from your personal mobile phone?	Multiple Choice	0000100001226185	Devices	◆ 0.563
Which resource should be used to determine acceptable products or procedures, at Questionmark for full disk encryption?	Multiple Choice	0000100001869120	Devices	◆ 0.609
Which of the following must you do if you want to sync your company emails to a personal mobile phone?	Multiple Response	6162783017786699	Devices	◆ 0.642

Questionmark
powered by Learnosity

How difficult should a question be?

- Differing p-values in a test are *normal* and *acceptable*
- Norm-referenced tests
  - A wide range of p-values helpful
- Criterion-referenced tests
  - p-values around the cut score helpful (e.g. often 0.6 to 0.8)

p-value	What it means
0	No-one gets the question right
< 0.25	Very hard question, most people get it wrong. Consider if should use.
0.25 to 0.9	Medium level – may be acceptable
> 0.9	Very easy question, almost everyone gets it right
1.0	Everyone gets the question right

Questionmark
powered by Learnosity

Common Reasons for Poor Item Difficulty

Too difficult

- Obscure content/has not been taught
- Poorly worded or confusing item
- Delivered at the end of a timed test
- Question scored wrongly
- Two choices that are both right

Too easy

- Well-known content
- Item has been exposed and shared
- Clue in item on what the right answer is
- Poor distractors (alternative choices)

Can I use very easy or very difficult questions?

Reasons to use difficult questions

- Needed by blueprint and only ones available
- Need to assess wide range of ability
- Job needs high performance (e.g. astronaut)

Reasons to use easy questions

- For retrieval practice
- Need to assess wide range of ability
- Build confidence / reduce anxiety
- Needed by blueprint and deemed important, even though nearly everyone knows the answer
- Compliance / health & safety questions - most people get it right, if someone gets it wrong, you want to flag

Example Question – What might the difficulty be?

What kind of animal is a dolphin?

a. Mammal
b. Aquatic
c. Cetacean
d. Fish-eating

Likely to be hard…
all choices are correct!

Example Question – What might the difficulty be?

A fertile area of desert in which the water table reaches the ground surface is called an

a. Oasis

b. Mirage

c. Water hole

d. Polder

Poll Question

What do you think the difficulty of this question might be?

Example Question – What might the difficulty be?

A fertile area of desert in which the water table reaches the ground surface is called an

[✓ a. Oasis]

b. Mirage

c. Water hole

d. Polder

Too easy.
Only one choice is grammatically correct.

Looking at Multiple Choice Distractors

Technique 2 of 3

Reminder on Multiple Choice Question Design

A multiple choice question has one correct answer and some incorrect choices (distractors)

- It's Important that:
  - Only one of the answers is right
  - Clues not given by grammar / length / style of answers
  - Each distractor answer is plausible

Question wording

○ Choice 1

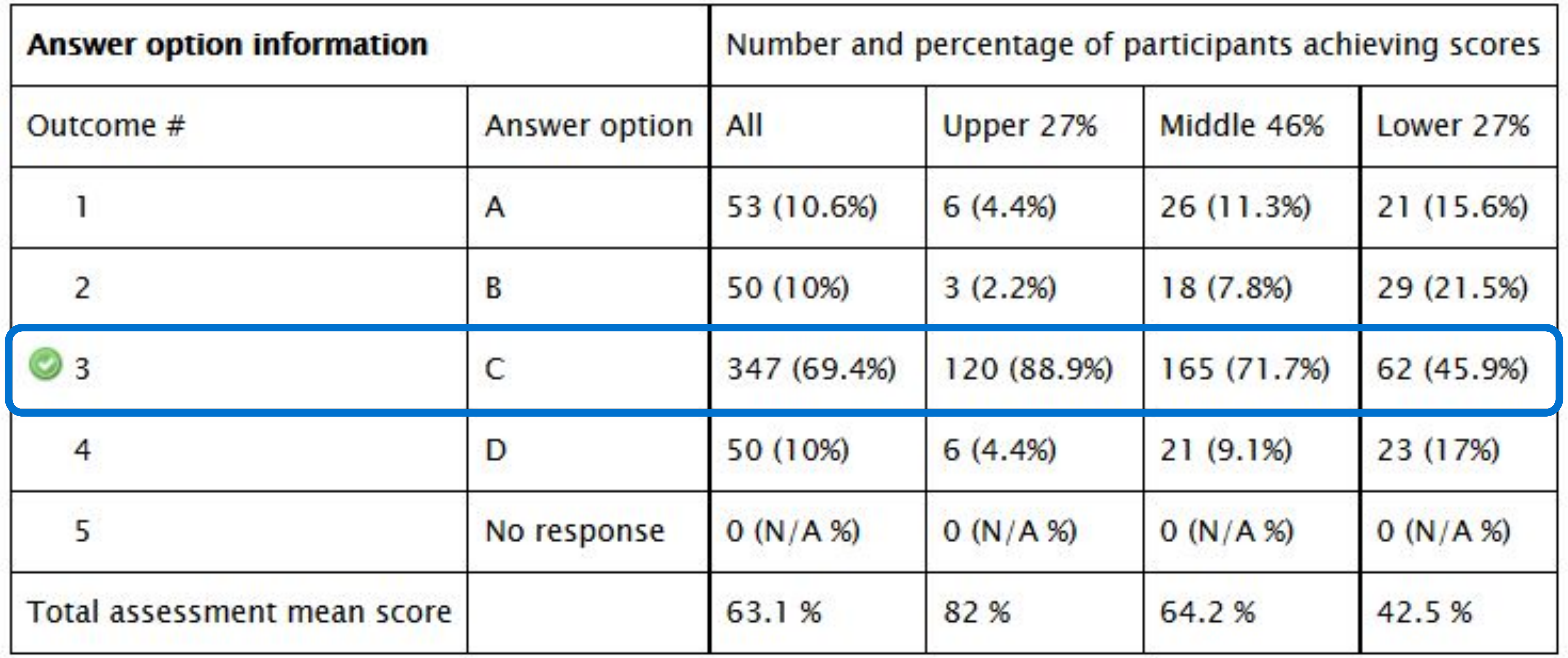
○ Choice 2

○ Choice 3

○ Choice 4

Answer Option Table

This table in Item Analysis report shows % of responses for each choice or outcome within each normative performance group: Upper, Middle, Lower



Answer option information		Number and percentage of participants achieving scores			
Outcome #	Answer option	All	Upper 27%	Middle 46%	Lower 27%
1	A	53 (10.6%)	6 (4.4%)	26 (11.3%)	21 (15.6%)
2	B	50 (10%)	3 (2.2%)	18 (7.8%)	29 (21.5%)
✓ 3	C	347 (69.4%)	120 (88.9%)	165 (71.7%)	62 (45.9%)
4	D	50 (10%)	6 (4.4%)	21 (9.1%)	23 (17%)
5	No response	0 (N/A %)	0 (N/A %)	0 (N/A %)	0 (N/A %)
Total assessment mean score		63.1 %	82 %	64.2 %	42.5 %

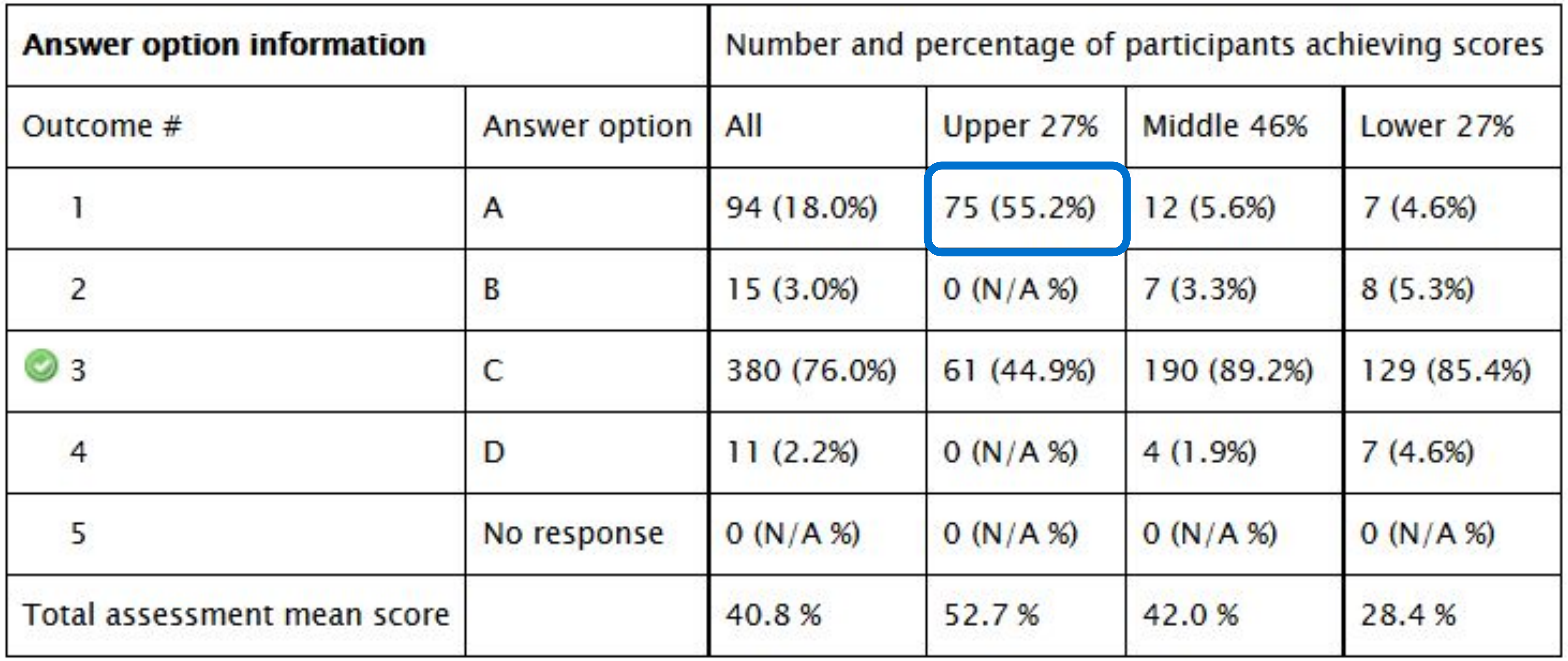
Uses of Answer Option Table

- Look at the # responses for each distractor to see if some distractors are not performing well.
  - If a distractor is not being selected, it could be a candidate for improvement
  - Good distractors often match common misconceptions or mistakes
- Look at the relative performance groups' percentage of responses to each option to identify:
  - Mis-scored items
  - Overlapping options
  - Differences in responses due to specialized knowledge

Uses of Answer Option Table

- Example where more than half of upper group selected a distractor

- C is correct, but some high-performing test-takers think A correct. Could be over-thinking, or may know of a case not considered in item-writing.



Answer option information		Number and percentage of participants achieving scores			
Outcome #	Answer option	All	Upper 27%	Middle 46%	Lower 27%
1	A	94 (18.0%)	75 (55.2%)	12 (5.6%)	7 (4.6%)
2	B	15 (3.0%)	0 (N/A %)	7 (3.3%)	8 (5.3%)
✓ 3	C	380 (76.0%)	61 (44.9%)	190 (89.2%)	129 (85.4%)
4	D	11 (2.2%)	0 (N/A %)	4 (1.9%)	7 (4.6%)
5	No response	0 (N/A %)	0 (N/A %)	0 (N/A %)	0 (N/A %)
Total assessment mean score		40.8 %	52.7 %	42.0 %	28.4 %

Do these questions need change?

Poll Question

Should you consider changing:

a. The question on the left
b. The question on the right
c. Both questions
d. Neither question

If 6 test-takers answer your 4-choice item like this:

a. 3
b. 3
c. 0
d. 0

600 test-takers answer your 4-choice item like this:

a. 295
b. 296
c. 8
d. 1

Do these questions need change?

If 6 test-takers answer your 4-choice item like this:

a. 3

b. 3

c. 0

d. 0

Do not change. Not enough data to make a decision.

600 test-takers answer your 4-choice item like this:

a. 295

b. 296

c. 8

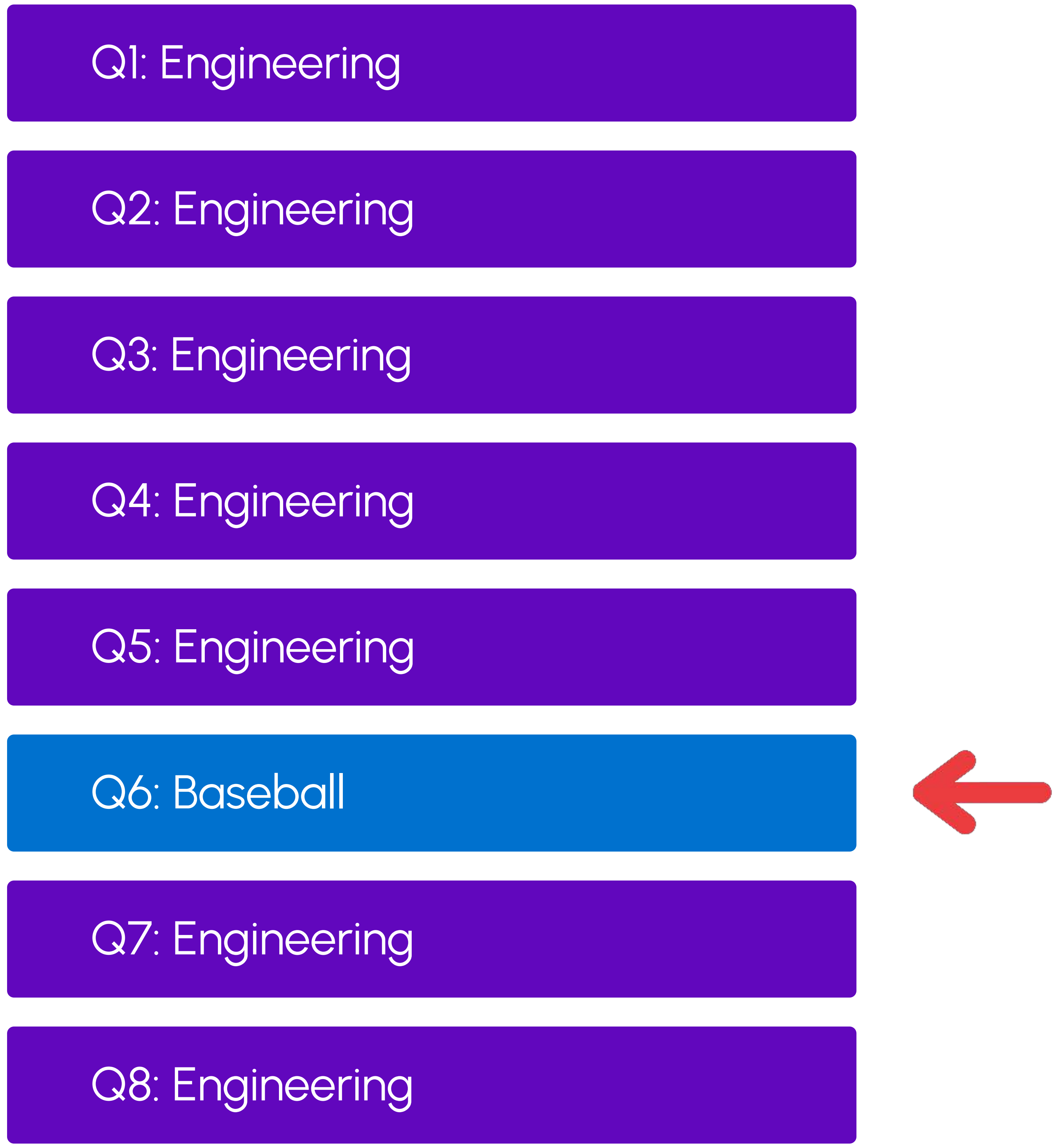
d. 1

Not enough people are choosing C and D. Should consider changing as makes question too easy to guess.

Item Discrimination: How well a Question Contributes to an Assessment Result

Technique 3 of 3

Consider this Test...



Q1: Engineering

Q2: Engineering

Q3: Engineering

Q4: Engineering

Q5: Engineering

Q6: Baseball

Q7: Engineering

Q8: Engineering

- What is the problem here?
  - Knowledge of baseball not relevant to an engineering test
  - Q6 reduces validity and reliability of the test
  - People who do well on Q6 may or may not do well on the other questions
  - How can we find questions like this in a real test?

Item Analysis Gives a Solution

- It's possible to look at
    - Test-takers who get a question correct
    - How well they do on the test as a whole
    - And work out a "correlation" between the two
- The Item Analysis report calculates "Item Discrimination" which is such a correlation
    - Number from -1.0 to +1.0
    - Compares how score for an item compares to score for the assessment
    - The higher it is, the better the item helps contribute to the assessment result

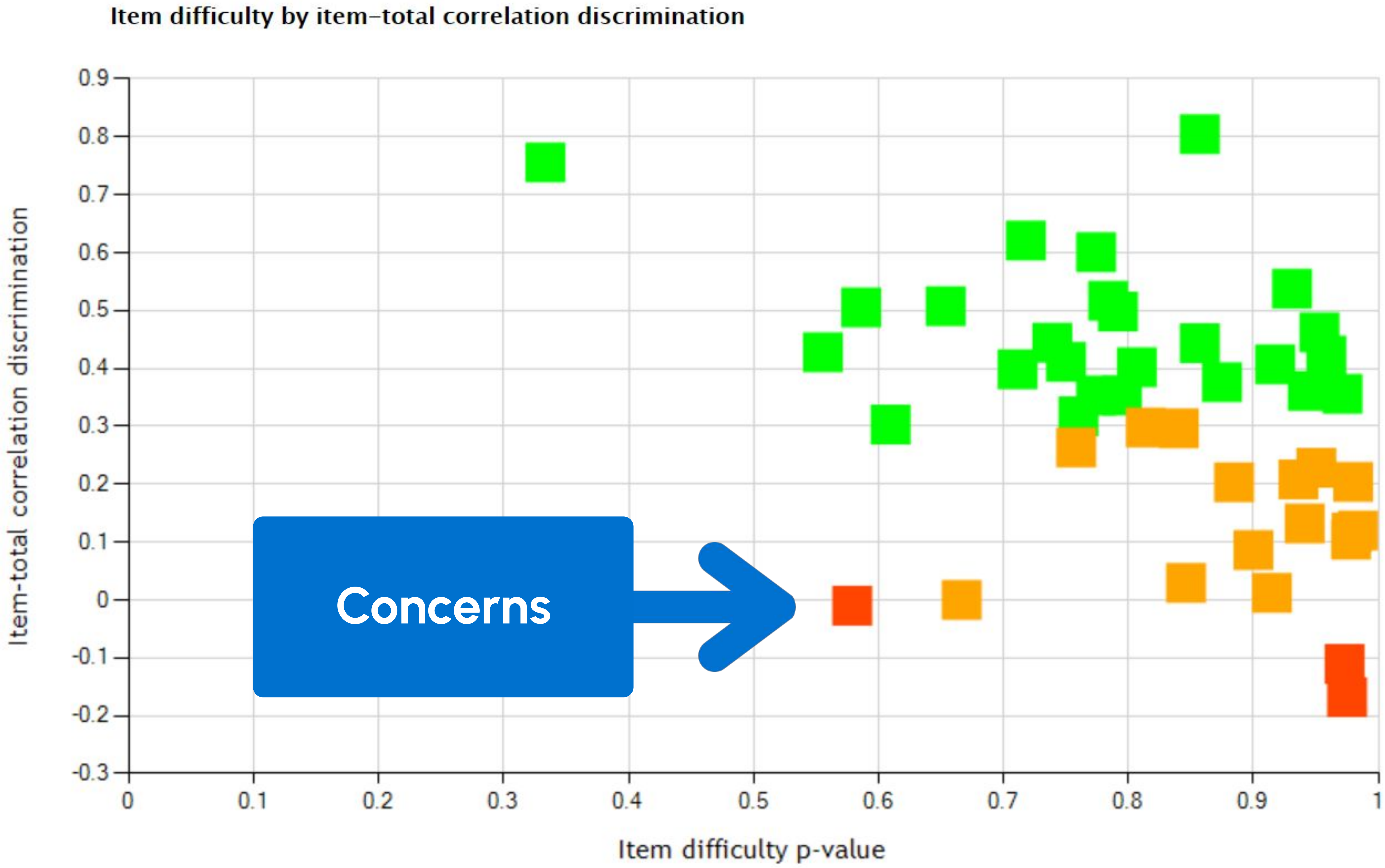
Item Discrimination

- Items with good discrimination improve assessment's ability to discriminate between test-takers of different ability levels.
- Item discrimination is influenced by p-value so expect lower values on very hard or very easy items.
- Items with low or negative discrimination may lower reliability of assessment or threaten validity (like the baseball example).

Possible Acceptable Range:

Discrimination = 0.20 – 1.00

Item Analytics Report Plots p-value vs Item Discrimination



Questionmark
powered by Learnosity

Some Reasons Why Discrimination May be Low

- Item is very easy
- Item is very hard
- Item correct answer is awry
- More than one correct answer
- Question is ambiguous or poorly written
- High-performing test-takers are overthinking the item
- Question is measuring different construct than other items
- Low sample size

Consider These Questions

Poll Question

	Item Difficulty	Item Discrimination
1	0.5	0.5
2	0.5	0.1
3	0.9	0.1
4	0.5	-0.1
5	0.6	0.2

Which questions need review?

a. None of them
b. Questions 2, 3, 4
c. Questions 2, 4
d. Question 1
e. All of them

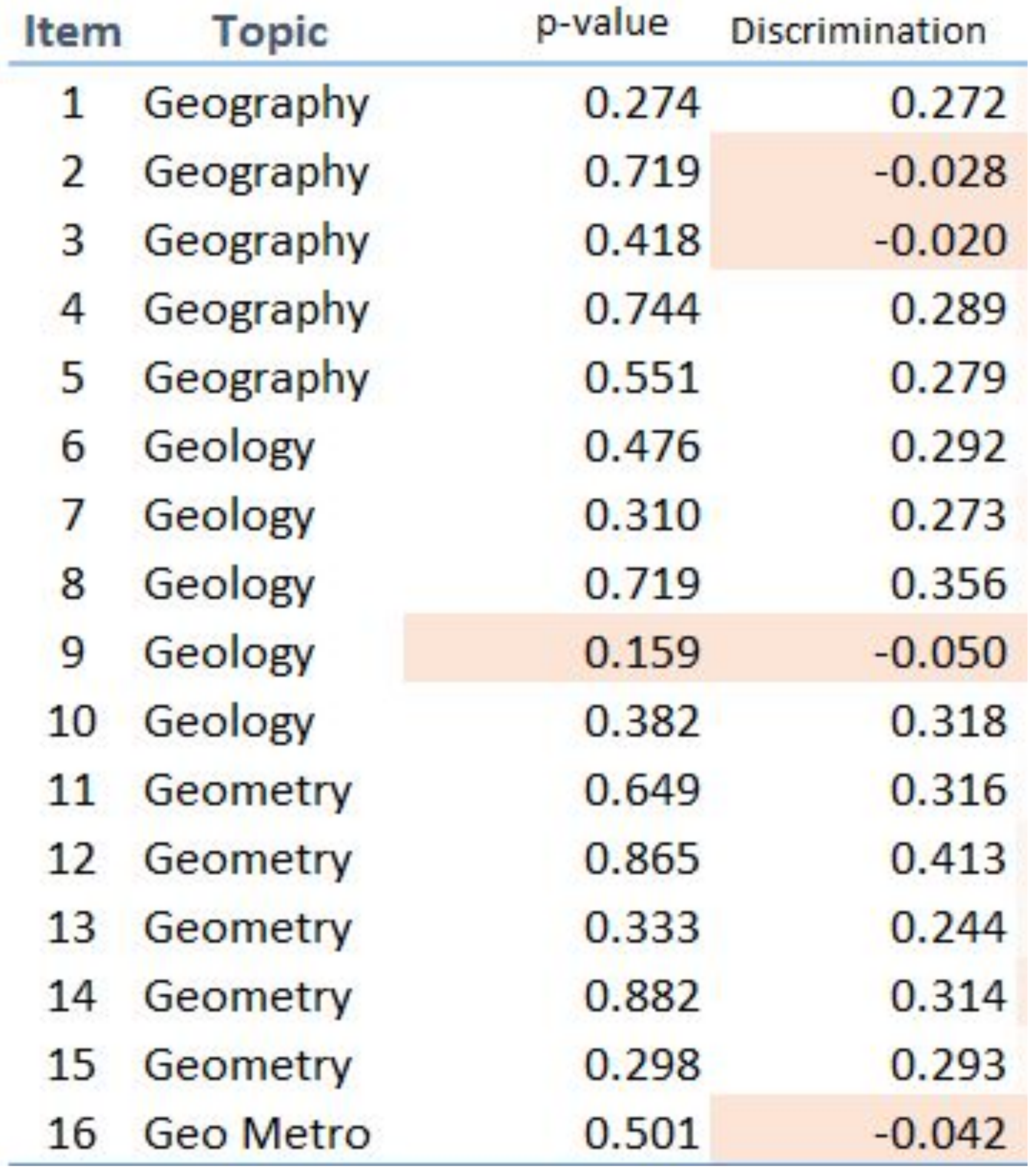
Item Analytics Report Plots p-value vs Item Discrimination

	Item Difficulty	Item Discrimination	What should you do?
1	0.5	0.5	Apparently good question
2	0.5	0.1	Marginal question - examine further
3	0.9	0.1	Very easy question. If acceptable, could consider keeping
4	0.5	-0.1	Poor question - probably needs change
5	0.6	0.2	Less discriminating than #1 but likely still good enough

Some Practical Examples

Scenario

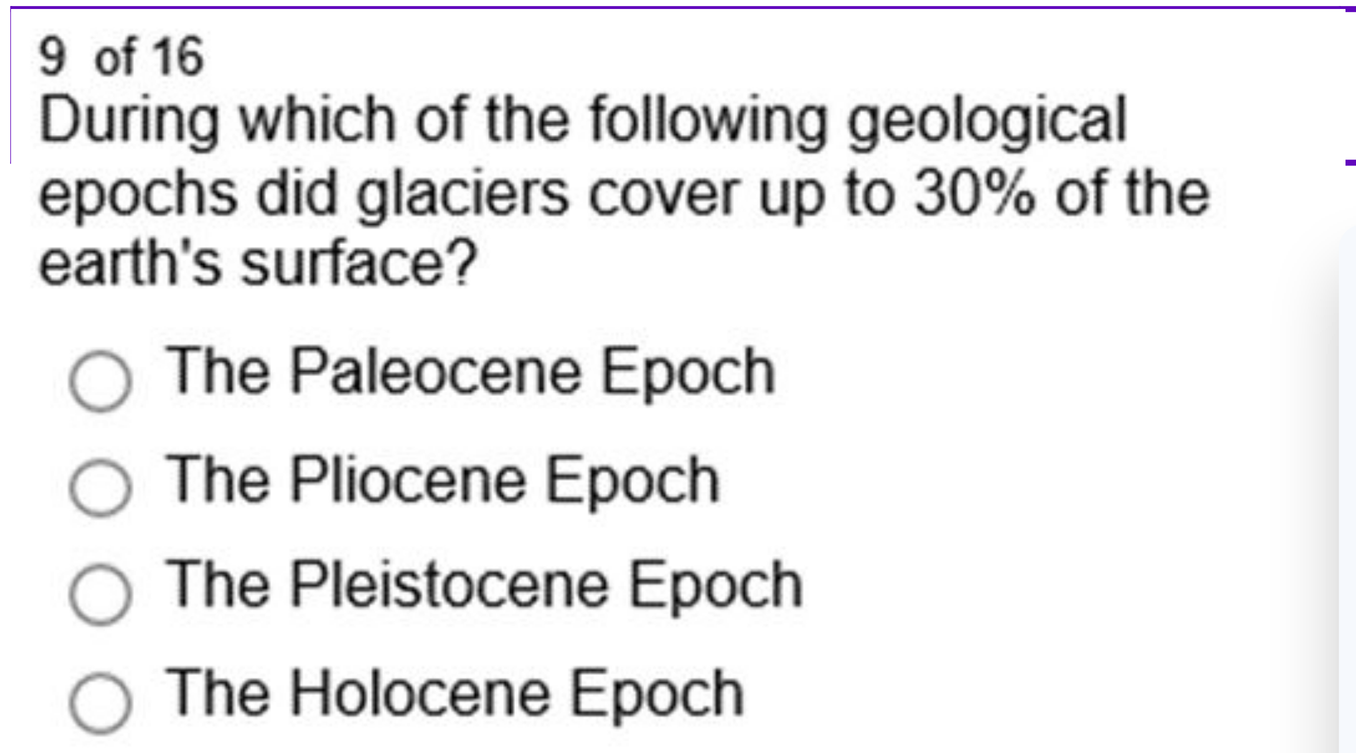
- Here are 16 questions from a larger test on topics beginning with "geo"
- Highlighted difficulty values outside of (0.25, 0.90)
- Highlighted discriminations below 0.20
- Let's look at the 4 problem items



Item	Topic	p-value	Discrimination
1	Geography	0.274	0.272
2	Geography	0.719	-0.028
3	Geography	0.418	-0.020
4	Geography	0.744	0.289
5	Geography	0.551	0.279
6	Geology	0.476	0.292
7	Geology	0.310	0.273
8	Geology	0.719	0.356
9	Geology	0.159	-0.050
10	Geology	0.382	0.318
11	Geometry	0.649	0.316
12	Geometry	0.865	0.413
13	Geometry	0.333	0.244
14	Geometry	0.882	0.314
15	Geometry	0.298	0.293
16	Geo Metro	0.501	-0.042

Question #1

Poll Question

Item Content	Item	Topic	Item Difficulty (p-value)
 9 of 16 During which of the following geological epochs did glaciers cover up to 30% of the earth's surface? ○ The Paleocene Epoch ○ The Pliocene Epoch ○ The Pleistocene Epoch ○ The Holocene Epoch	9	Geology	0.159

Why does this item have a low p-value?

a. It's mis-keyed

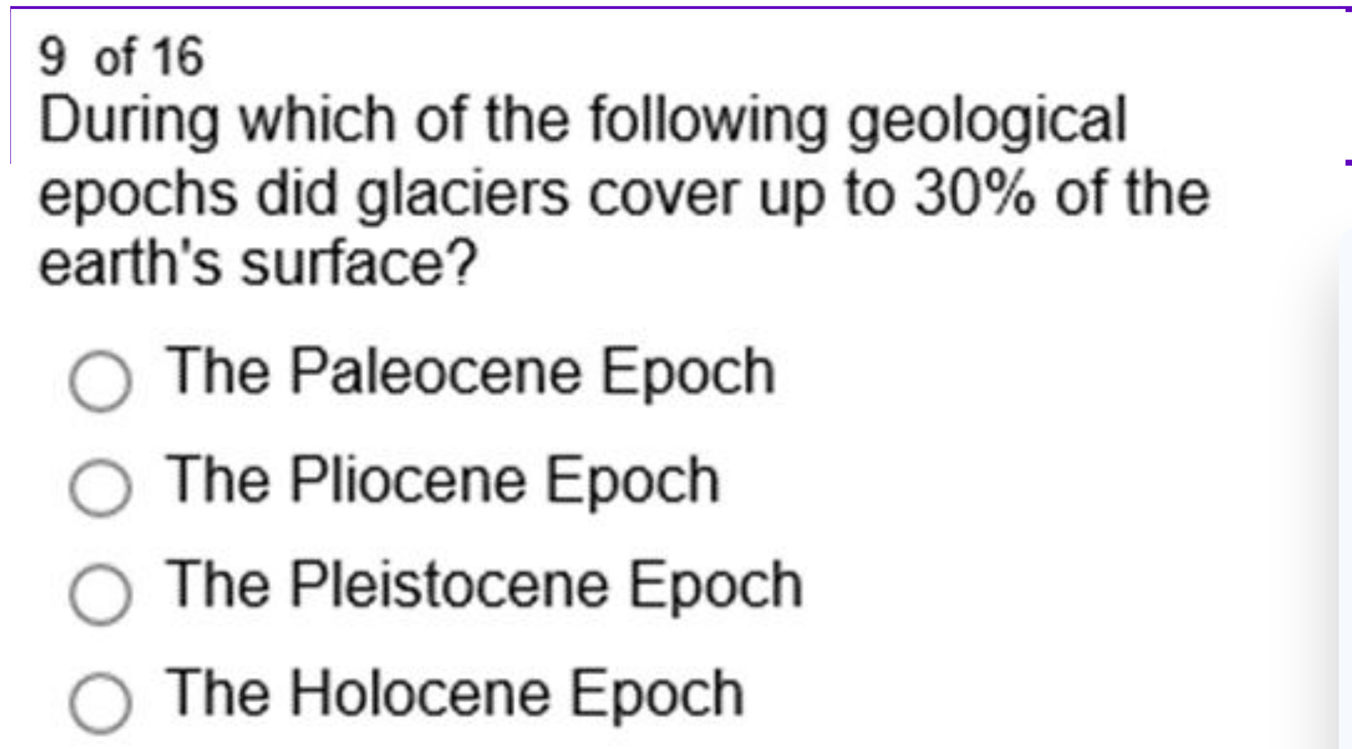
b. It's poorly worded/confusing

c. It has two keys/overlapping options

d. Other

✔	Answer option	All	Upper 27%	Middle 46%	Lower 27%
	The Paleocene Epoch	304 (30.4%)	39 (14.4%)	148 (32.2%)	117 (43.3%)
✔	The Pliocene Epoch	159 (15.9%)	67 (24.8%)	62 (13.5%)	30 (11.1%)
	The Pleistocene Epoch	284 (28.4%)	130 (48.1%)	139 (30.2%)	15 (5.6%)
	The Holocene Epoch	253 (25.3%)	34 (12.6%)	111 (24.1%)	108 (40%)

Answer #1

Item Content	Item	Topic	Item Difficulty (p-value)
 9 of 16 During which of the following geological epochs did glaciers cover up to 30% of the earth's surface? ○ The Paleocene Epoch ○ The Pliocene Epoch ○ The Pleistocene Epoch ○ The Holocene Epoch	9	Geology	0.159

Why does this item have a low p-value?

a. It's mis-keyed

b. It's poorly worded/confusing

c. It has two keys/overlapping options

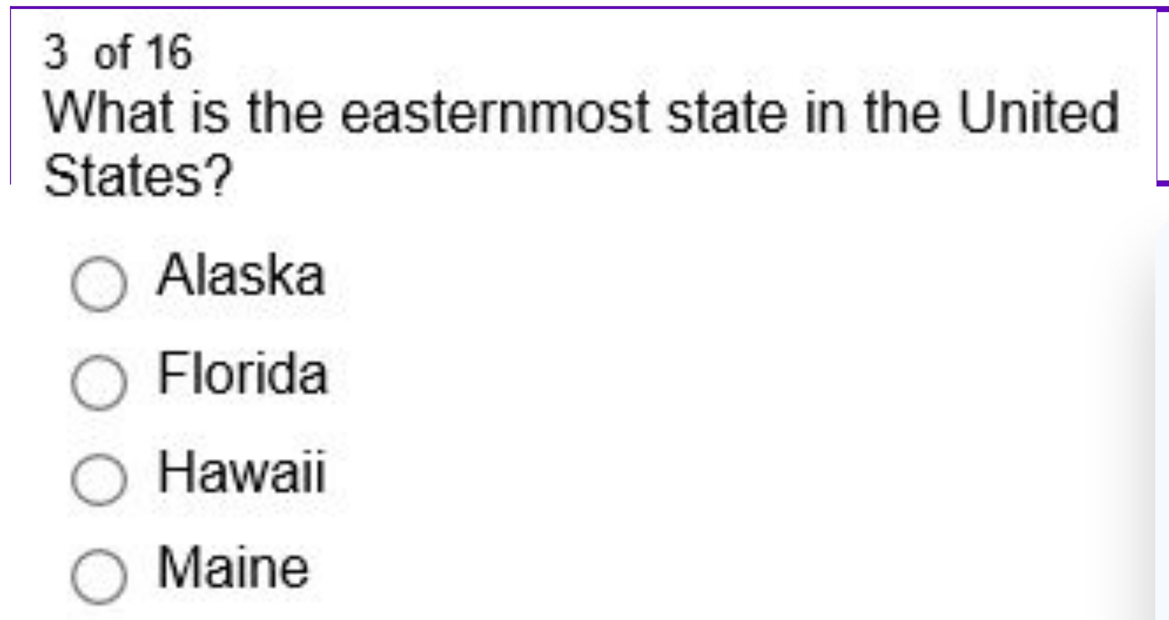
d. Other

Correct response is “The Pleistocene Epoch.”

✔	Answer option	All	Upper 27%	Middle 46%	Lower 27%
	The Paleocene Epoch	304 (30.4%)	39 (14.4%)	148 (32.2%)	117 (43.3%)
✔	The Pliocene Epoch	159 (15.9%)	67 (24.8%)	62 (13.5%)	30 (11.1%)
	The Pleistocene Epoch	284 (28.4%)	130 (48.1%)	139 (30.2%)	15 (5.6%)
	The Holocene Epoch	253 (25.3%)	34 (12.6%)	111 (24.1%)	108 (40%)

Question #2

Poll Question

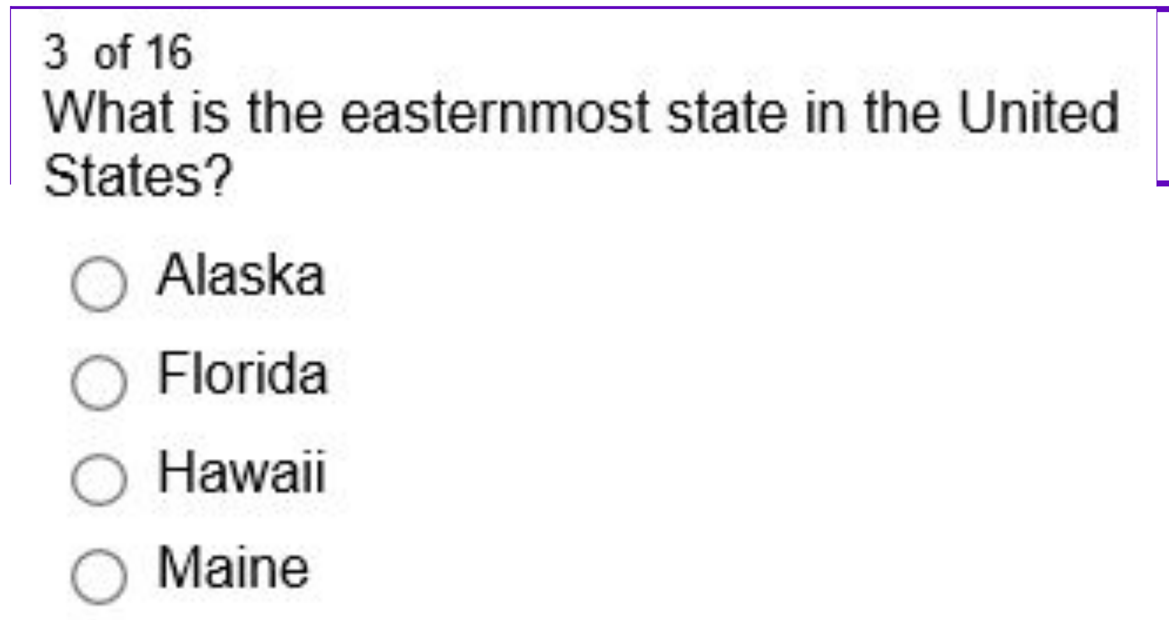
Item Content	Item	Topic	Item Difficulty (p-value)	Item Discrimination
 3 of 16 What is the easternmost state in the United States? ○ Alaska ○ Florida ○ Hawaii ○ Maine	3	Geography	0.418	-0.020

Why do you think this item has a low discrimination value?

a. It's mis-keyed
b. It's poorly worded/confusing
c. High-performers are overthinking it
d. It is measuring a different construct
e. Other

✔	Answer option	All	Upper 27%	Middle 46%	Lower 27%
	Alaska	278 (27.8%)	117 (43.3%)	91 (19.8%)	70 (25.9%)
	Florida	146 (14.6%)	17 (6.3%)	63 (13.7%)	66 (24.4%)
	Hawaii	158 (15.8%)	17 (6.3%)	71 (15.4%)	70 (25.9%)
✔	Maine	418 (41.8%)	119 (44.1%)	235 (51.1%)	64 (23.7%)

Answer #2

Item Content	Item	Topic	Item Difficulty (p-value)	Item Discrimination
 3 of 16 What is the easternmost state in the United States? ○ Alaska ○ Florida ○ Hawaii ○ Maine	3	Geography	0.418	-0.020

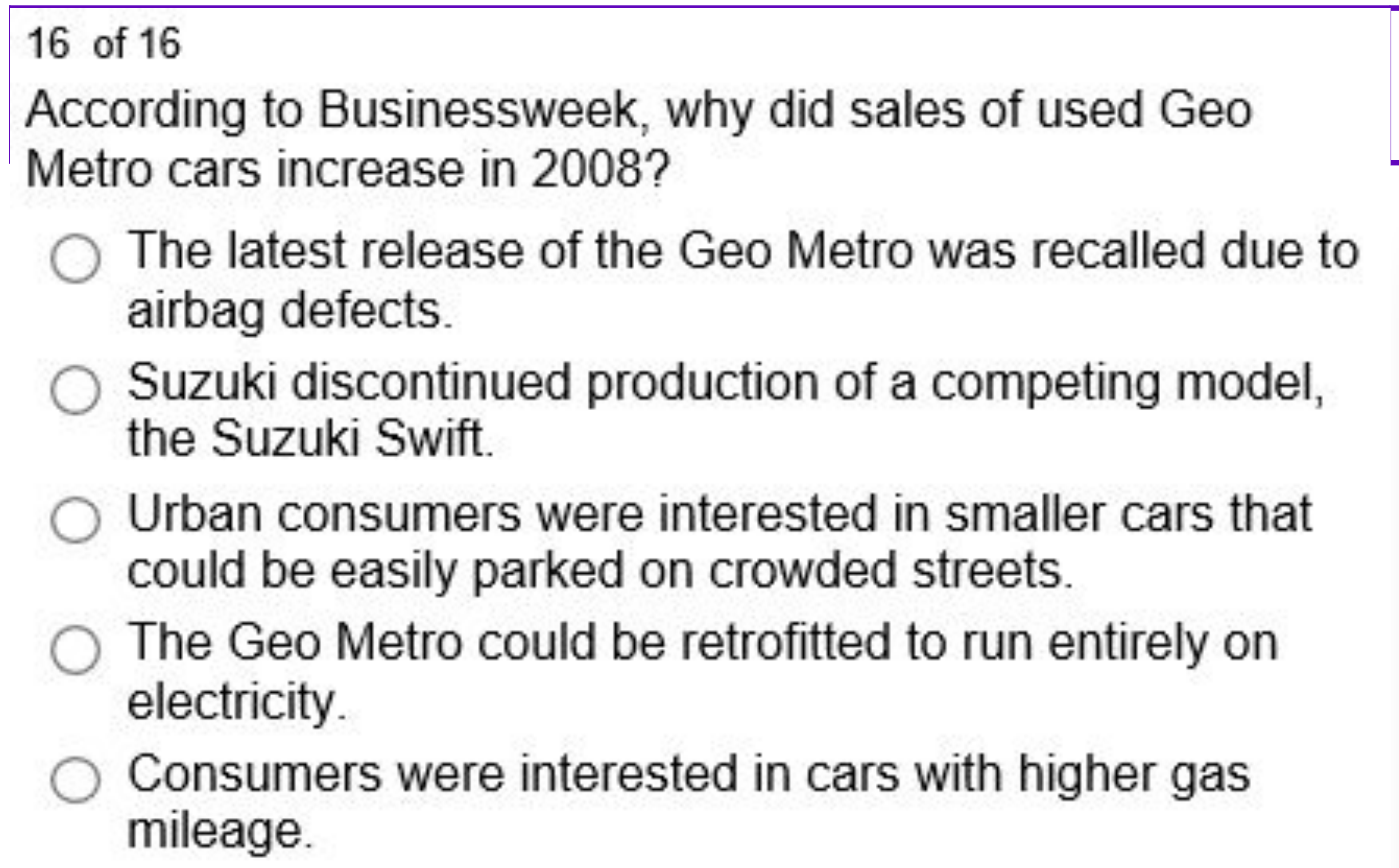
You can see that many highly performing test-takers choose Alaska

Likely reason: test-takers with specialized knowledge identified Alaska is the easternmost state because Pochnoi Point is past the 180° longitude. Technically mis-keyed. (A)

✓	Answer option	All	Upper 27%	Middle 46%	Lower 27%
	Alaska	278 (27.8%)	117 (43.3%)	91 (19.8%)	70 (25.9%)
	Florida	146 (14.6%)	17 (6.3%)	63 (13.7%)	66 (24.4%)
	Hawaii	158 (15.8%)	17 (6.3%)	71 (15.4%)	70 (25.9%)
✓	Maine	418 (41.8%)	119 (44.1%)	235 (51.1%)	64 (23.7%)

Question #3

Poll Question

Item Content	Item	Topic	Item Difficulty (p-value)	Item Discrimination
	16	Geo Metro	0.501	-0.042

16 of 16

According to Businessweek, why did sales of used Geo Metro cars increase in 2008?

○ The latest release of the Geo Metro was recalled due to airbag defects.

○ Suzuki discontinued production of a competing model, the Suzuki Swift.

○ Urban consumers were interested in smaller cars that could be easily parked on crowded streets.

○ The Geo Metro could be retrofitted to run entirely on electricity.

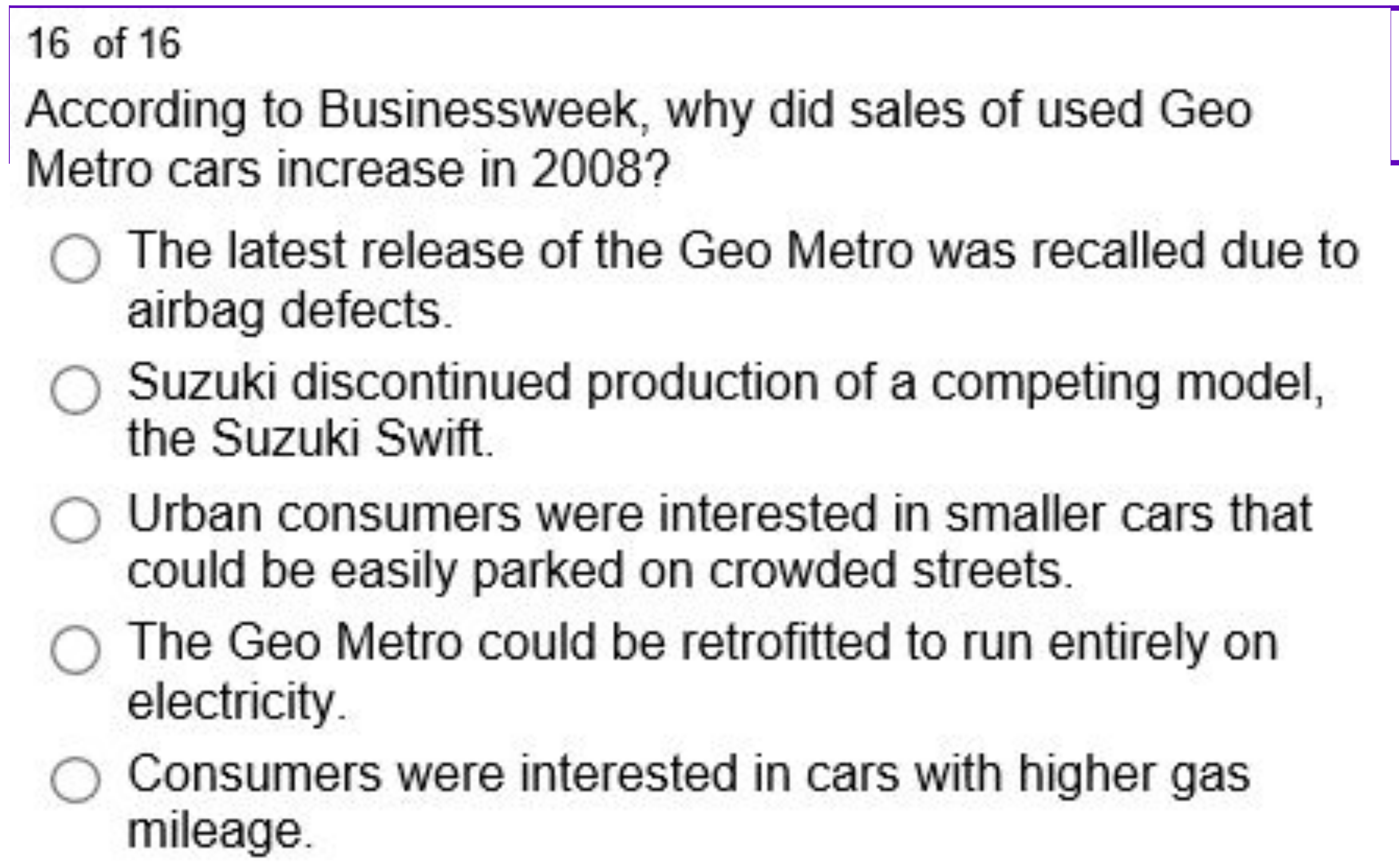
○ Consumers were interested in cars with higher gas mileage.

Why do you think this item has a low discrimination value?

a. It's mis-keyed
b. It's poorly worded/confusing
c. High-performers are overthinking it
d. It is measuring a different construct
e. Other

✓	Answer option	All	Upper 27%	Middle 46%	Lower 27%
	The latest release of the Geo Metro . . .	136 (13.6%)	36 (13.3%)	65 (14.1%)	35 (13.0%)
	Suzuki discontinued production of a . . .	138 (13.8%)	46 (17.0%)	55 (12.0%)	37 (13.7%)
	Urban consumers were interested in . . .	113 (11.3%)	33 (12.2%)	39 (8.5%)	41 (15.2%)
	The Geo Metro could be retrofitted. . .	112 (11.2%)	33 (12.2%)	51 (11.1%)	28 (10.4%)
✓	Consumers were interested in cars . . .	501 (50.1%)	122 (45.2%)	250 (54.3%)	129 (47.8%)

Answer #3

Item Content	Item	Topic	Item Difficulty (p-value)	Item Discrimination
	16	Geo Metro	0.501	-0.042

16 of 16

According to Businessweek, why did sales of used Geo Metro cars increase in 2008?

○ The latest release of the Geo Metro was recalled due to airbag defects.

○ Suzuki discontinued production of a competing model, the Suzuki Swift.

○ Urban consumers were interested in smaller cars that could be easily parked on crowded streets.

○ The Geo Metro could be retrofitted to run entirely on electricity.

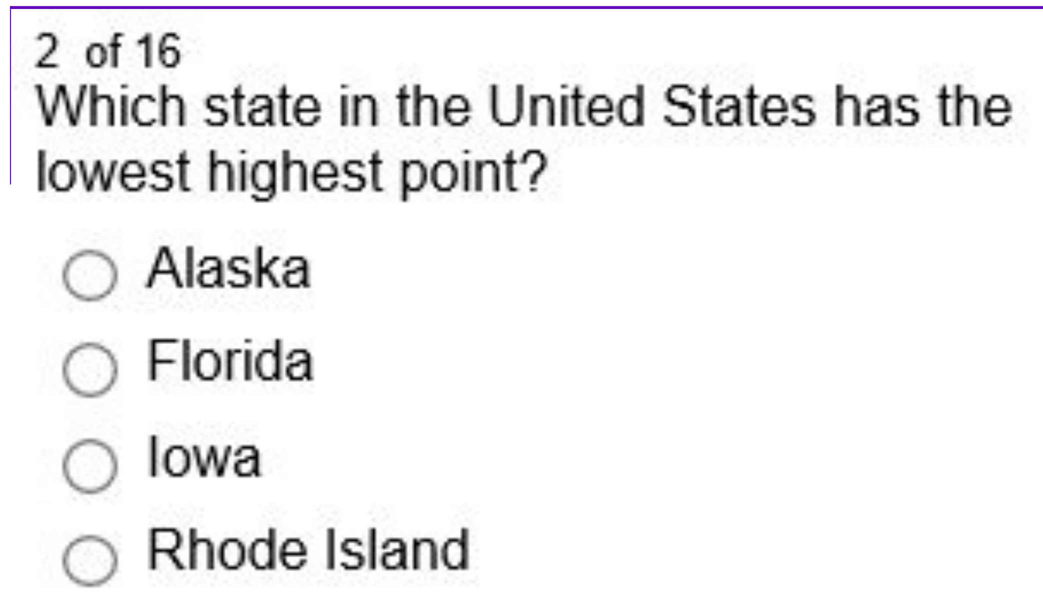
○ Consumers were interested in cars with higher gas mileage.

The most likely reason is that this is measuring a different construct to other questions (D)

✔	Answer option	All	Upper 27%	Middle 46%	Lower 27%
	The latest release of the Geo Metro . . .	136 (13.6%)	36 (13.3%)	65 (14.1%)	35 (13.0%)
	Suzuki discontinued production of a . . .	138 (13.8%)	46 (17.0%)	55 (12.0%)	37 (13.7%)
	Urban consumers were interested in . . .	113 (11.3%)	33 (12.2%)	39 (8.5%)	41 (15.2%)
	The Geo Metro could be retrofitted. . .	112 (11.2%)	33 (12.2%)	51 (11.1%)	28 (10.4%)
✔	Consumers were interested in cars . . .	501 (50.1%)	122 (45.2%)	250 (54.3%)	129 (47.8%)

Question #4

Poll Question

Item Content	Item	Topic	Item Difficulty (p-value)	Item Discrimination
 2 of 16 Which state in the United States has the lowest highest point? ○ Alaska ○ Florida ○ Iowa ○ Rhode Island	2	Geography	0.719	-0.028

Why do you think this item has a low discrimination value?

a. It's mis-keyed

b. It's poorly worded/confusing

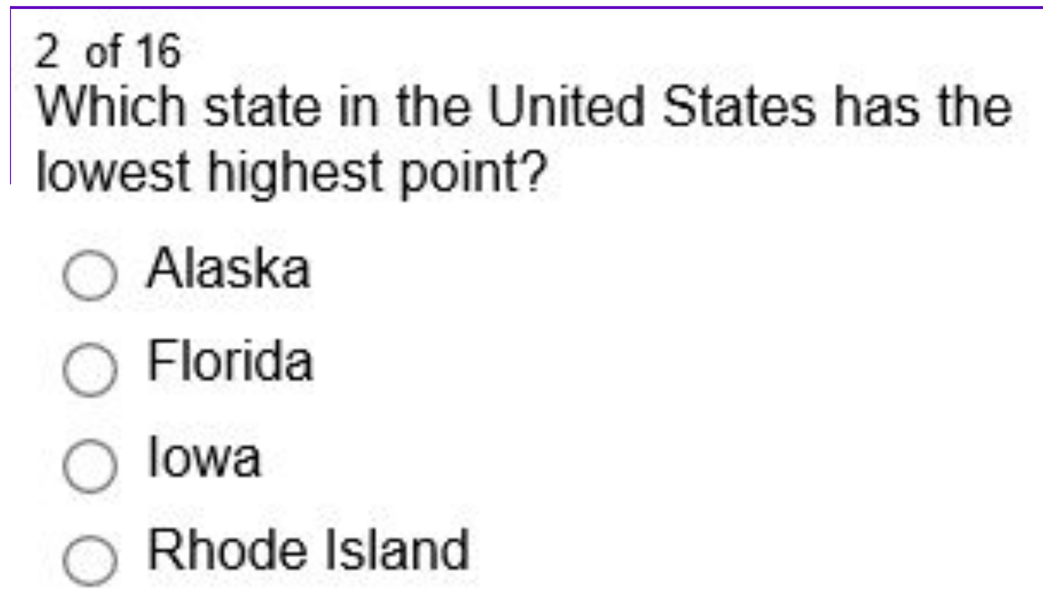
c. High-performers are overthinking it

d. It is measuring a different construct

e. Other

✔	Answer option	All	Upper 27%	Middle 46%	Lower 27%
	Alaska	101 (10.1%)	14 (5.2%)	55 (12%)	32 (11.9%)
✔	Florida	719 (71.9%)	212 (78.5%)	328 (71.3%)	179 (66.3%)
	Iowa	90 (9.0%)	19 (7.0%)	38 (8.3%)	33 (12.2%)
	Rhode Island	90 (9.0%)	25 (9.3%)	39 (8.5%)	26 (9.6%)

Answer #4

Item Content	Item	Topic	Item Difficulty (p-value)	Item Discrimination
 2 of 16 Which state in the United States has the lowest highest point? ○ Alaska ○ Florida ○ Iowa ○ Rhode Island	2	Geography	0.719	-0.028

No identifiable issues with content, but low discrimination means we should probably drop in favor of better questions (E)

✔	Answer option	All	Upper 27%	Middle 46%	Lower 27%
	Alaska	101 (10.1%)	14 (5.2%)	55 (12%)	32 (11.9%)
✔	Florida	719 (71.9%)	212 (78.5%)	328 (71.3%)	179 (66.3%)
	Iowa	90 (9.0%)	19 (7.0%)	38 (8.3%)	33 (12.2%)
	Rhode Island	90 (9.0%)	25 (9.3%)	39 (8.5%)	26 (9.6%)

Review of Decisions to Remove

Item	Description	Item Difficulty	Item Discrimination	Rationale for Decision to Remove
9	Glaciers	0.159	-0.050	Mis-keyed. Correct response is "The Pleistocene Epoch."
3	Easternmost State	0.418	-0.020	Test-takers with specialized knowledge identified that Alaska is the easternmost state because Pochnoi Point is past the 180° longitude. Technically mis-keyed.
16	Geo Metro Sales (2008)	0.501	-0.042	Item scores do not correlate well with total scores. It may be that knowledge of Geo Metro sales is not related to the construct that test is designed to measure.
2	Lowest Highest Point	0.719	-0.028	No identifiable issues with content, but low discrimination means we should probably drop in favor of better questions

Reminder of Key Vocabulary

Distractor

Wrong choice in a multiple choice question

Item

Question

Item Discrimination

Number between -1.0 and 1.0 which shows item's correlation to test score, also called item-total correlation

p-value

Number between 0.0 and 1.0 which shows item's difficulty

Reliability

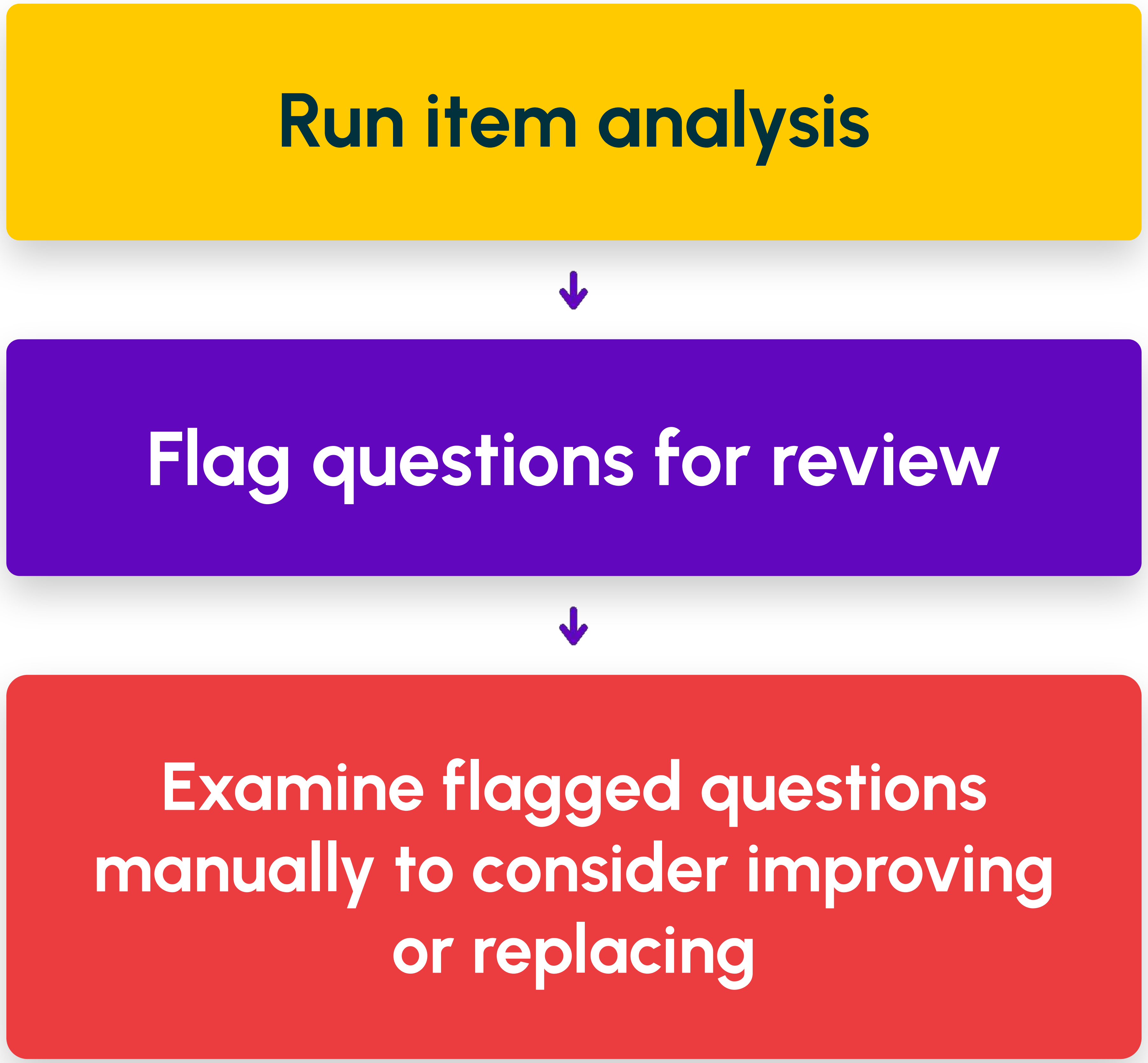
How consistent the assessment is

Validity

Whether the assessment measures what it seeks to measure

Summary

- Item analysis flags questions for review
  - Difficulty too high or low
  - Distractors awry
  - Discrimination awry
- Useful tool to help review how your questions work in practice
- There is lots more to learn, but what covered today a useful start



Run item analysis

Flag questions for review

Examine flagged questions manually to consider improving or replacing

Questions?

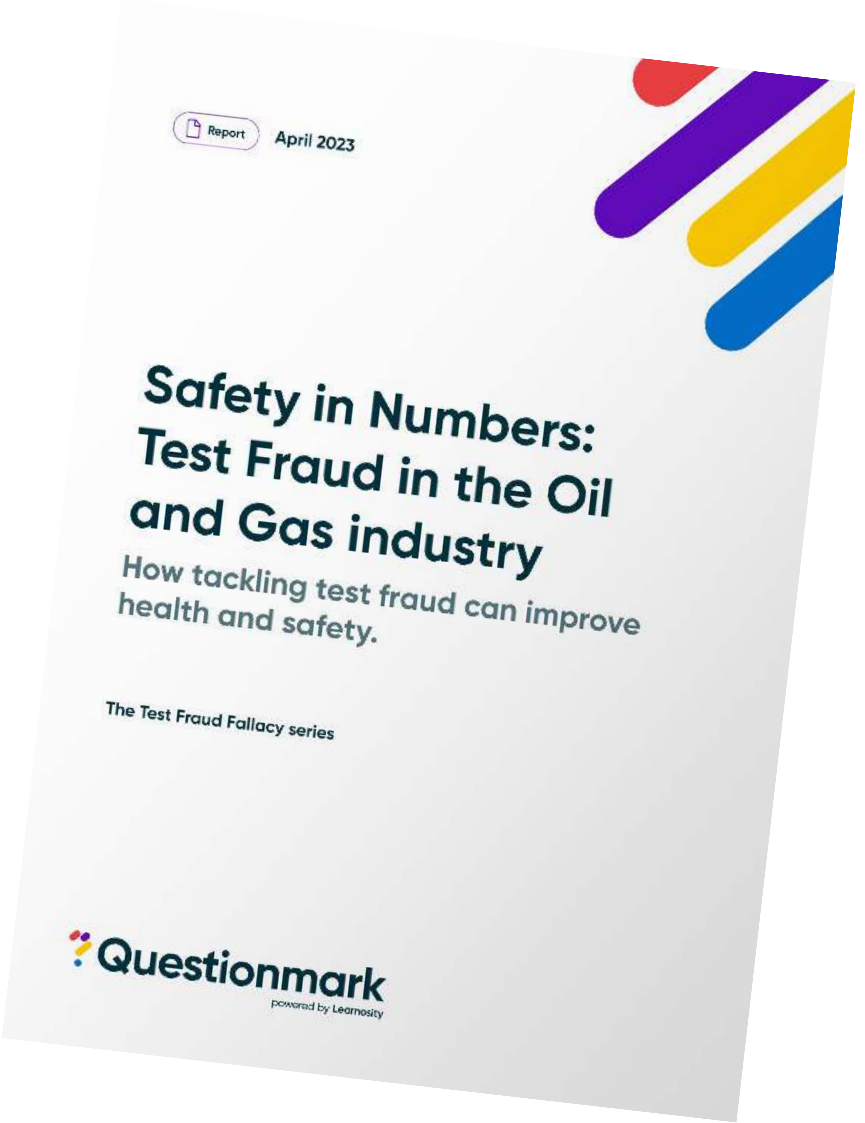
Resources

Questionmark's resources

Check out our website to find the latest:

- Reports
- Webinars
- Blog Articles
- Podcast Episodes
- And more!

Visit questionmark.com



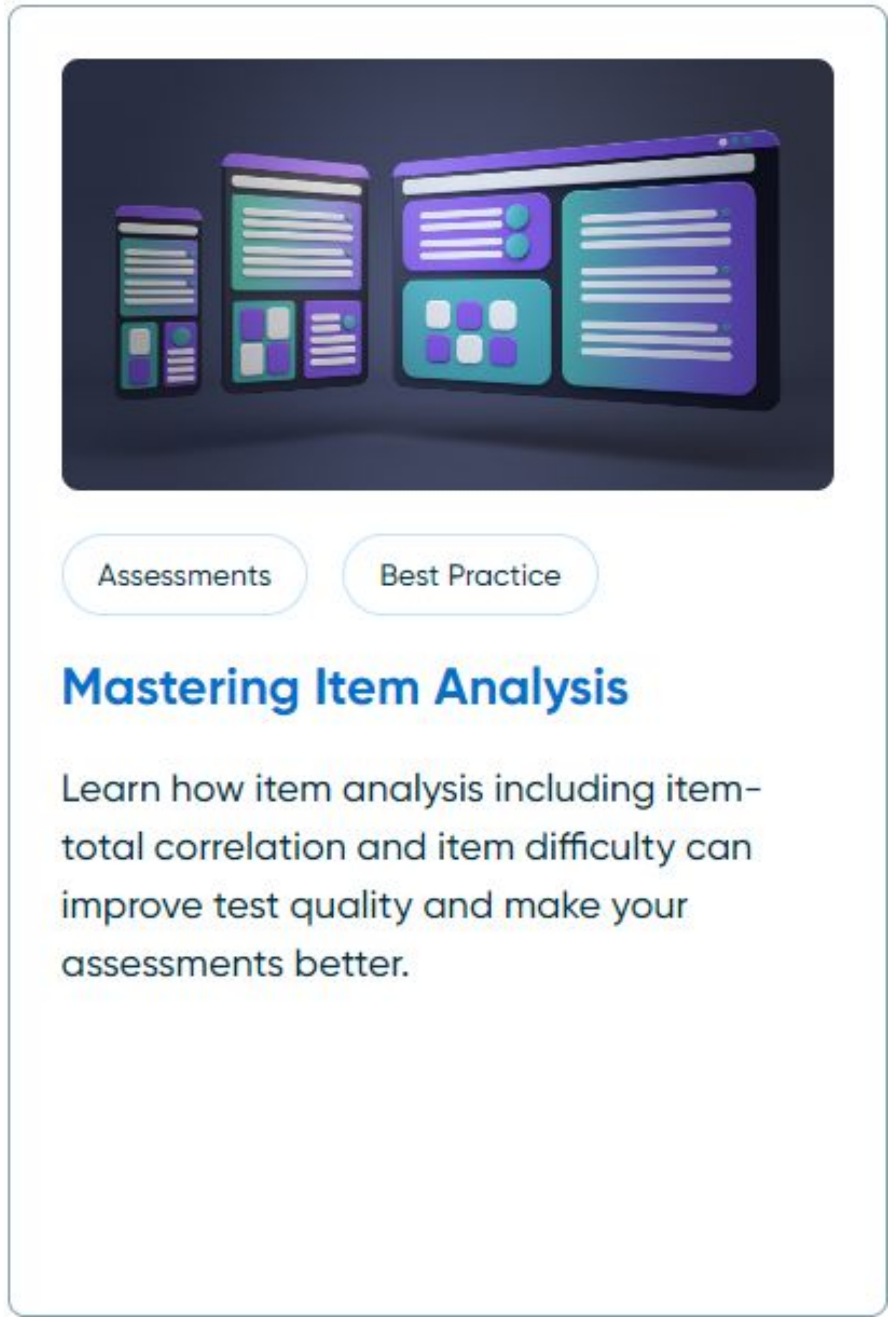
Report April 2023

Safety in Numbers: Test Fraud in the Oil and Gas industry

How tackling test fraud can improve health and safety.

The Test Fraud Fallacy series

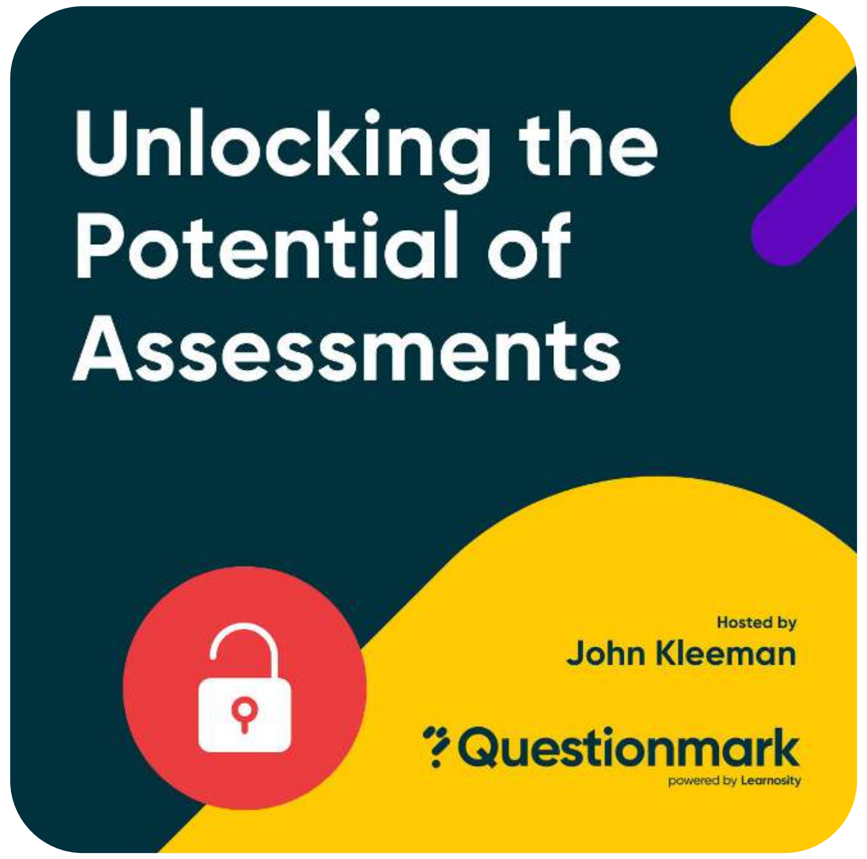
Questionmark
powered by Learnosity



Assessments Best Practice

Mastering Item Analysis

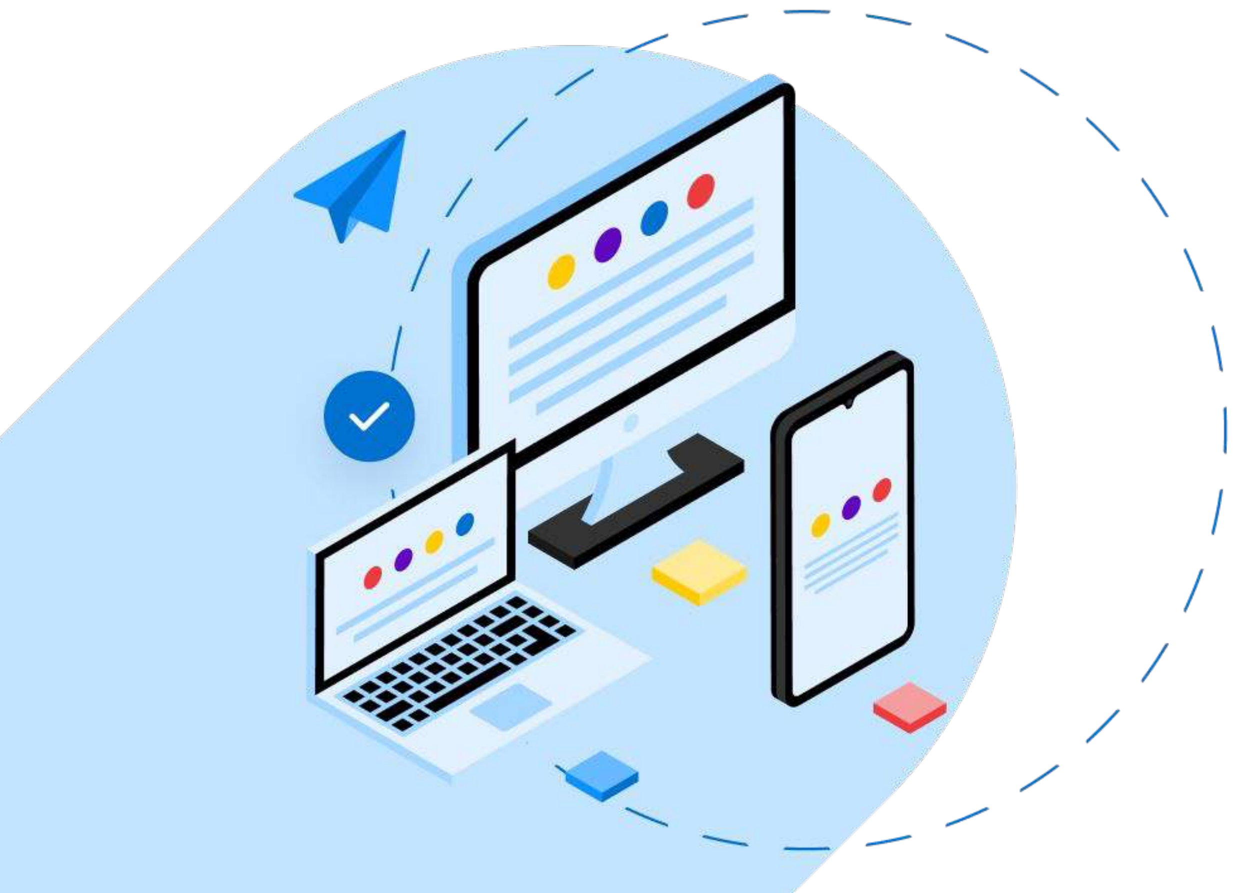
Learn how item analysis including item-total correlation and item difficulty can improve test quality and make your assessments better.



Unlocking the Potential of Assessments

Hosted by
John Kleeman

Questionmark
powered by Learnosity



Questionmark

powered by Learnosity

questionmark.com