An Overview of the GRE Subject Test in Mathematics; ETS Test Development Process

Walt Jiménez
Senior Director
ETS Global Higher Education

11/05/2020

Educational Testing Service (ETS)

- The world's largest private not-for-profit educational testing and measurement organization.
- Each year, we develop, administer and score more than 50 million tests — including the TOEFL®, TOEIC®, GRE® and Praxis® tests — in more than 180 countries at more than 9,000 locations.
- Our mission is to advance quality and equity in education by providing fair and valid assessments, research and related services.
- Our products and services measure knowledge and skills, promote learning and performance, and support education and professional development for all people worldwide.

Graduate Record Examinations® (GRE ®)

- The GRE tests were created to provide an objective lens through which all graduate school applicants can be compared, regardless of their background.

- GRE General Test
  - Verbal Reasoning
  - Quantitative Reasoning
  - Analytical Writing
- GRE Subject Tests
  - Biology
  - Chemistry
  - Literature in English
  - Mathematics
  - Physics
  - Psychology

GRE Mathematics Test—Overview

- The test consists of approximately 66 multiple-choice questions drawn from courses commonly offered at the undergraduate level.
- Test duration is 170 minutes.
- Approximately 50 percent of the questions involve calculus and its applications — subject matter that is assumed to be common to the backgrounds of almost all mathematics majors.
- About 25 percent of the questions in the test are in elementary algebra, linear algebra, abstract algebra, and number theory.
- The remaining questions deal with other areas of mathematics currently studied by undergraduates in many institutions.

ETS Test Development Process

Step 1: Defining Objectives

- Who will take the test and for what purpose?
- What skills and/or areas of knowledge should be tested?
- How should test takers be able to use their knowledge?
- What kinds of questions should be included? How many of each kind?
- How long should the test be?
- How difficult should the test be?

Step 2: Item Development Committees

- Responsibilities of these item development committees may include:
  - defining test objectives and specifications
  - helping ensure test questions are unbiased
  - determining test format (e.g., multiple-choice, essay, constructed-response, etc.)
  - considering supplemental test materials
  - reviewing test questions, or test items, written by ETS staff
  - writing test questions

Step 3: Writing and Reviewing Questions

- Each test question — written by ETS staff or item development committees — undergoes numerous reviews and revisions to ensure it is as clear as possible, that it has only one correct answer among the options provided on the test and that it conforms to the style rules used throughout the test.

Item (Question) Development Process

1. Item author (including stimulus)
2. Content Review 1
3. Content Review 2
4. Content Review 3 (Optional)
5. Fairness Review
6. Edit Review
7. External Review (Optional)
8. Owner Resolution
9. Final Format
10. Final Content Review

Step 4: The Pretest

- After the questions have been written and reviewed, many are pretested with a sample group similar to the population to be tested. The results enable test developers to determine:
  - the difficulty of each question
  - if questions are ambiguous or misleading
  - if questions should be revised or eliminated
  - if incorrect alternative answers should be revised or replaced

Step 5: Detecting and Removing Unfair Questions

- To meet the stringent **ETS Standards for Quality and Fairness** guidelines, Trained reviewers carefully inspect each individual test question, the test as a whole and any descriptive or preparatory materials to ensure that language, symbols, words, phrases and content generally regarded as sexist, racist or otherwise inappropriate or offensive to any subgroup of the test-taking population are eliminated.
- ETS statisticians also identify questions on which two groups of test takers who have demonstrated similar knowledge or skills perform differently on the test through a process called **Differential Item Functioning (DIF)**. If one group performs consistently better than another on a particular question, that question receives additional scrutiny and may be deemed biased or unsatisfactory. Note: If people in different groups actually differ in their average levels of relevant knowledge or skills, a fair test question will reflect those differences.

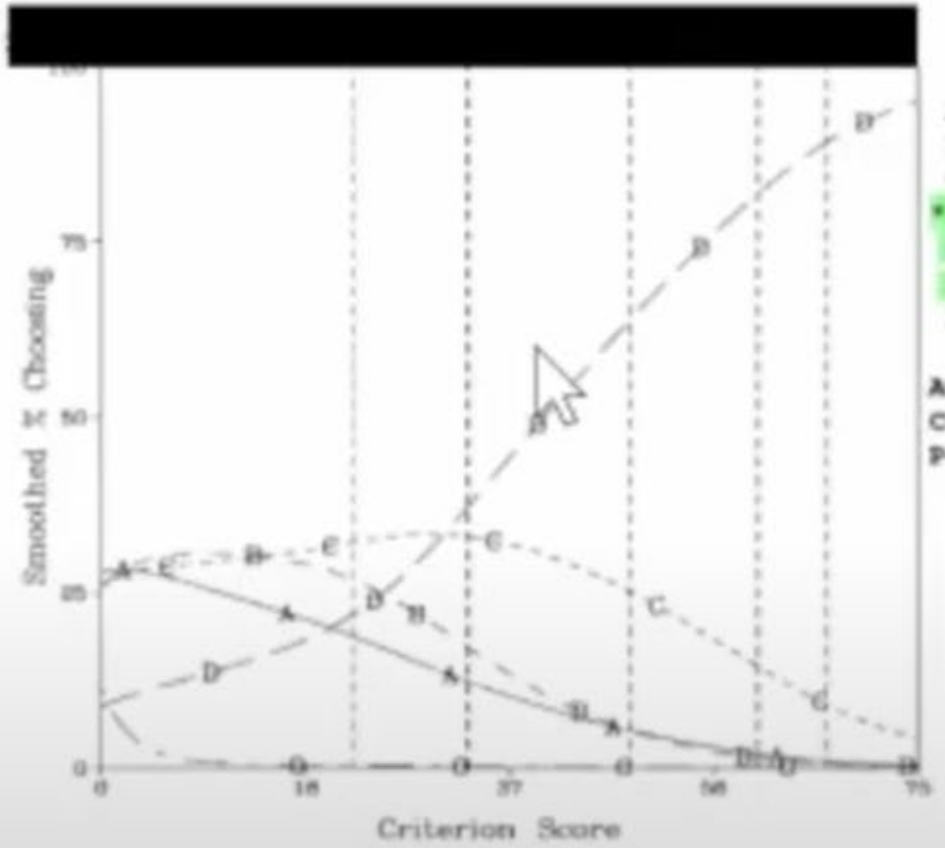
Step 6: Assembling the Test

- After the test is assembled, it is reviewed by other specialists, committee members and sometimes other outside experts. Each reviewer answers all questions independently and submits a list of correct answers to the test developers. The lists are compared with the ETS answer keys to verify that the intended answer is, indeed, the correct answer. Any discrepancies are resolved before the test is published.

Step 7: Making Sure — Even After the Test is Administered — that the Test Questions are Functioning Properly

- Even after the test has been administered, statisticians and test developers review to make sure that test questions are working as intended. Before final scoring takes place, each question undergoes preliminary statistical analysis and results are reviewed question by question. If a problem is detected, such as the identification of a misleading answer to a question, corrective action, such as not scoring the question, is taken before final scoring and score reporting takes place.
- Tests are also reviewed for reliability. Performance on one version of the test should reasonably predict performance on any other version of the test. If reliability is high, results will be similar no matter which version a test taker completes.

Sample "Good" Item (4-option MCQ)

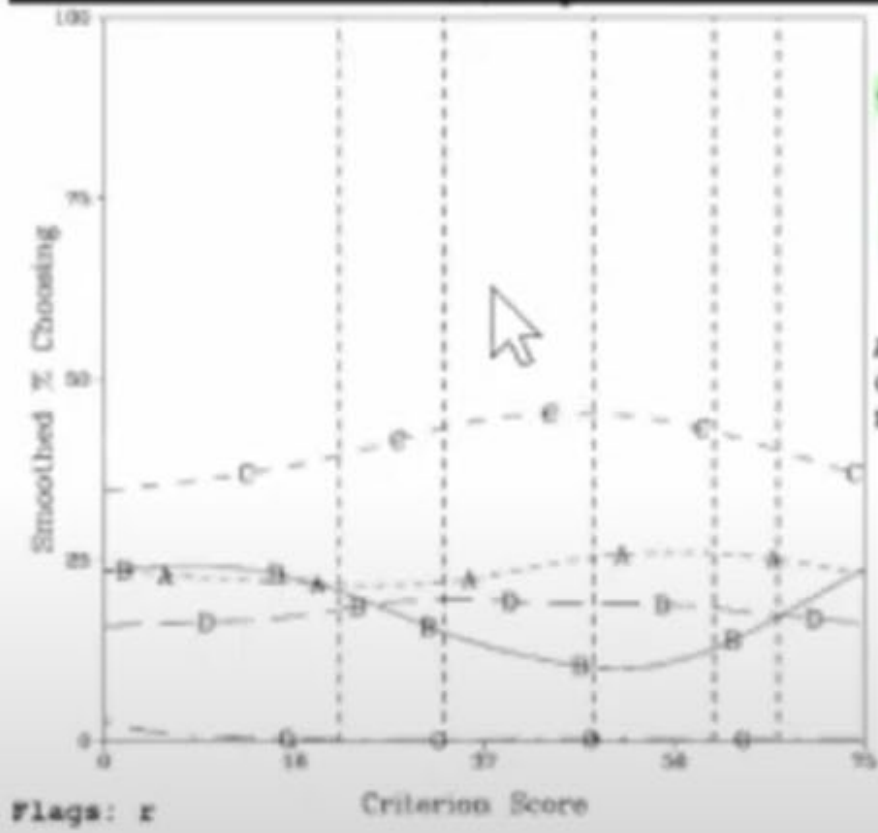


	Multiple Choice, 4 choice				
			Criterion		Top
	N	%Tot	Mean	SD	20%
A	32191	7.8	32.6	13.0	0.7
B	40635	9.9	30.7	11.4	0.5
C	93940	22.8	40.4	14.1	7.5
*D	245051	59.4	53.6	13.2	91.2
Omt	707	0.2	35.4	16.7	0.1
NR	0				
Rch	412524	100.0	46.7	15.9	

Average Item Score	0.59
Correlation with Crit.	0.62
Percent Reached	100.00

Sample "Bad" Item (4-option MCQ)

[redacted]



Flags: r

Multiple Choice. 4 choice					
			Criterion		Top
	N	%Tot	Mean	SD	20%
A	5060	24.0	47.3	15.5	24.6
*B	3200	15.2	45.3	18.4	20.1
C	8889	42.2	46.5	15.2	38.4
D	3877	18.4	46.2	15.4	16.8
Omt	47	0.2	43.8	18.5	0.2
NR	0				
Rch	21073	100.0	46.5	15.9	

Average Item Score	0.15
Correlation with Crit.	-0.04
Percent Reached	100.00

[redacted]