

Human-in-the-loop automated item creation for complex items in language assessments

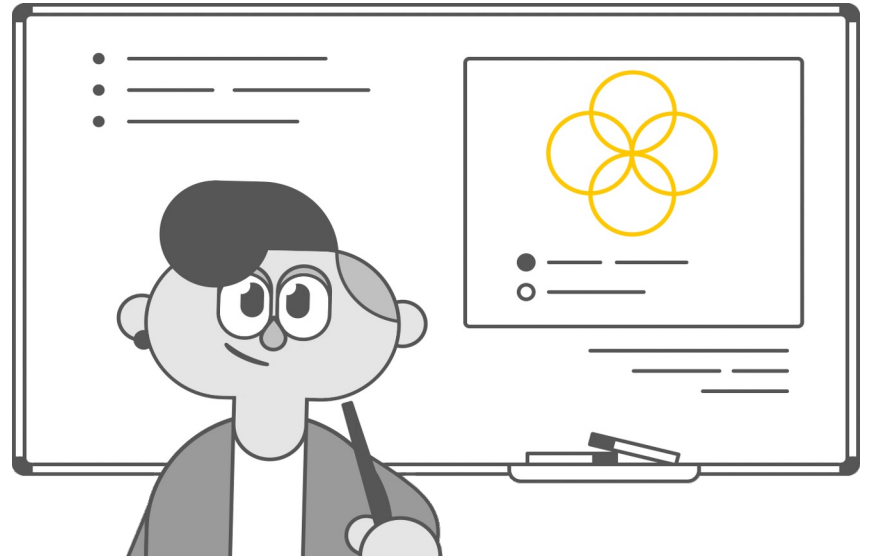
Geoffrey T. LaFlair, Yigal Attali, Andrew Runge,
Kevin Yancey, Sarah Goodwin, Yena Park,
Alina A. von Davier

March 21, 2023



The plan

- Overview of digital-first assessment
- Interactive reading
- Automated item generation
- Human review
- Pilot results



A digital-first assessment

Born digital

- Designed with computational models, AI-algorithms, and tech capabilities in mind
- Theory-based & evidence-centered
- Valid & reliable

Integrated

- Fluid infrastructure
- Integrative theoretical frameworks
- Integrated (automatic) tools & databases

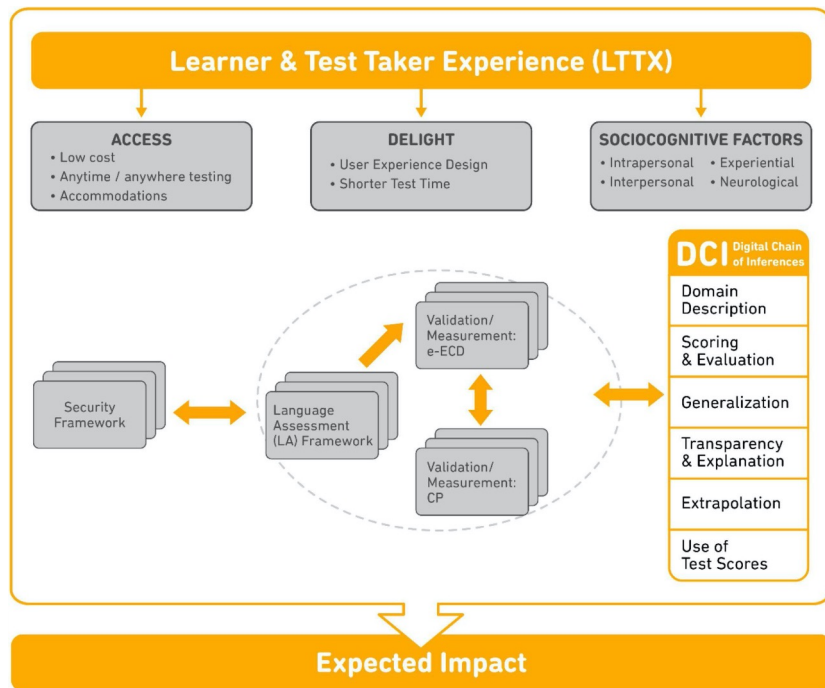
Superior test-taker experience (TTX)

- Available on demand, anytime, anywhere
- Online testing with remote proctoring



Digital-first design

How can we use technology to enhance how we measure language ability?



Integrate digital affordances

- Task design
- Test delivery
- Task interaction
- Task scoring
- Security
- Psychometrics

The complex items: Interactive reading

Interactive Reading

The image displays three overlapping screenshots of a digital reading test interface. The top-left screenshot shows a timer at 7:21 and a 'QUIT TEST' button. The middle-left screenshot shows a timer at 6:49 and a 'QUIT TEST' button. The bottom-right screenshot shows a timer at 6:07 and a 'QUIT TEST' button. Each screenshot features a 'PASSAGE' section with text and a question section with multiple-choice options. The text in the passages is about interior design and the role of interior designers. The questions are about the meaning of words and the purpose of the text.

7:21 for the next 6 questions

QUIT TEST

PASSAGE

There is no single, unifying definition of the term "interior design", but most people understand it as referring to how owners and designers of buildings and permanent residences can best use the space to create a great living experience. Interior designers have been changing the way people think of and interact with their living spaces for more than a hundred years. The first interior designers were responsible for dressing wealthy homes. They were in charge of keeping up the interior design of their employer's home, so the owner always had a beautiful residence for entertaining. They learned how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms.

Select the best option for each missing word

2 analogous to how owners and designers c

3 and permanent residences can

6:49 for the next 5 questions

QUIT TEST

PASSAGE

There is no single, unifying definition of the term "interior design", but most people understand it as referring to how owners and designers of buildings and permanent residences can best use the space to create a great living experience. Interior designers have been changing the way people think of and interact with their living spaces for more than a hundred years. The first interior designers were responsible for dressing wealthy homes. They were in charge of keeping up the interior design of their employer's home, so the owner always had a beautiful residence for entertaining. They learned how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms.

6:07 for the next 3 questions

QUIT TEST

PASSAGE

There is no single, unifying definition of the term "interior design", but most people understand it as referring to how owners and designers of buildings and permanent residences can best use the space to create a great living experience. Interior designers have been changing the way people think of and interact with their living spaces for more than a hundred years. The first interior designers were responsible for dressing wealthy homes. They were in charge of keeping up the interior design of their employer's home, so the owner always had a beautiful residence for entertaining. They learned how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms.

Click and drag text to highlight the answer to the question below

What did early interior designers learn?

how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms

NEXT

Targeted reading skills

- Vocabulary knowledge
- Discourse/text completion
- Identify information (x2)
- Read to understand
- Read to use information

Technology & task interaction

The top screenshot displays a test interface with a timer at 7:21 and a 'QUIT TEST' button. The question asks to select the best option for a missing word in a passage. The dropdown menu is set to 'definition'.

The bottom screenshot displays a test interface with a timer at 6:49 and a 'QUIT TEST' button. It shows a reading passage about interior design and a multiple-choice question. The third option is selected.

Top Screenshot:

7:21 for the next 6 questions QUIT TEST

PASSAGE

There is no single, unifying **1 definition** of the term "interior design", but most people understand it as

1 definition

2 analogous to how owners and designers of

3

4

experience the way living spaces

Bottom Screenshot:

6:49 for the next 5 questions QUIT TEST

PASSAGE

There is no single, unifying definition of the term "interior design", but most people understand it as referring to how owners and designers of buildings and permanent residences can best use the space to create a great living experience. Interior designers have been changing the way people think of and interact with their living spaces for more than a hundred years.

The first interior designers were responsible for dressing wealthy homes.

They were in charge of keeping up the interior design of their employer's home, so the owner always had a beautiful residence for entertaining. They learned how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms.

Select the best sentence to fill in the blank in the passage

☐ The natural materials and tools available in each region influenced design choices.

☐ High rises became standard, as did indoor toilets and indoor plumbing.

☒ The first interior designers were responsible for dressing wealthy homes.

☐ Wooden furniture with modern accents is featured in one room after another.

NEXT

- TTs "see" how their choice fits into the broader context
- Staged reveal of the text (and some feedback)

Technology & task interaction

6:07 for the next 3 questions QUIT TEST

PASSAGE

There is no single, unifying definition of the term "interior design", but most people understand it as referring to how owners and designers of buildings and permanent residences can best use the space to create a great living experience. Interior designers have been changing the way people think of and interact with their living spaces for more than a hundred years. The first interior designers were responsible for dressing wealthy homes. They were in charge of keeping up the interior design of their employer's home, so the owner always had a beautiful residence for entertaining. They learned how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms

Click and drag text to highlight the answer to the question below

What did early interior designers learn?

how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms

NEXT

- Unique-stem question (x2)
- Scan for the information
- Highlight the answer
- Authentic interaction with the text

Identify the idea & Title the passage

5:58 for the next 2 questions

QUIT TEST

PASSAGE

There is no single, unifying definition of the term "interior design", but most people understand it as referring to how owners and designers of buildings can best use the space to create a great living experience. Interior designers have been changing the way people think of and interact with their living spaces for more than a hundred years. The first interior designers were responsible for dressing wealthy homes. They were in charge of keeping up the interior design of their employer's home, so the owner always had a beautiful residence for entertaining. They learned how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms.

Select the idea that is expressed in the passage

5:47 for this question

QUIT TEST

PASSAGE

There is no single, unifying definition of the term "interior design", but most people understand it as referring to how owners and designers of buildings and permanent residences can best use the space to create a great living experience. Interior designers have been changing the way people think of and interact with their living spaces for more than a hundred years. The first interior designers were responsible for dressing wealthy homes. They were in charge of keeping up the interior design of their employer's home, so the owner always had a beautiful residence for entertaining. They learned how to produce different colors that would complement each other, as well as how to hide ugly aspects of the house and cleverly disguise dark, dreary rooms.

Select the best title for the passage

- ☒ The History of Interior Design
- ☐ Designing The Family Room
- ☐ Interior Design Trends
- ☐ Origins of the Traditional Color Palette

NEXT

- Choose the idea that is discussed in the passage
- Use the information in the text to select the best title

How it's made: Automated item generation

The old approach to AIG

- **Item models**

- $[9] + ? = [6]$
- How many pieces of [fruit] will you have if you cut [5] whole [fruits] into [thirds]?

- **Advantages**

- Opportunity for a construct-driven approach to item development
- Expands the potential number of items

- **Serious limitations**

- Only simple item types
- Semi-automatic costly process, depends on experts creating models

New approach: Language modeling

Two related goals

- Understand how probable some text is
 - How likely is the word “dog” in a given sentence?
- Be able to generate plausible new text
 - Fill in the blank “I walked my ____”
 - How likely is each word in the corpus, given its context?

Richness of context is key

- Previous: N-gram language models → context is 5-10 words
- Current: GPT-3 → context 4,000 tokens, GPT-4 → context is 32,000 tokens

Technology and content generation

Generative Pretrained Transformer¹

- Extremely powerful language generation model
- Most ML models require hundreds or thousands of examples of task.
- GPT-3* excels at few-shot learning - can mirror style, format, and content given extremely few examples

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}
GPT-3 Small	125M	12	768	12	64
GPT-3 Medium	350M	24	1024	16	64
GPT-3 Large	760M	24	1536	16	96
GPT-3 XL	1.3B	24	2048	24	128
GPT-3 2.7B	2.7B	32	2560	32	80
GPT-3 6.7B	6.7B	32	4096	32	128
GPT-3 13B	13.0B	40	5140	40	128
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128



Passages & titles

Goal

Generate self-contained passages in a range of domains - especially at an academic level

Method

Generate passages and titles conditioned by other academic content and a specified topic using prototypical examples



Question generation

Common-stem questions

- Main point of passage

Example: X-rays are used in various branches of engineering and science, including medicine and communications.

Unique-stem questions

- Identify information

Example: How did Wilhelm Röntgen's discovery change the way that medical imaging was done?

MC option generation

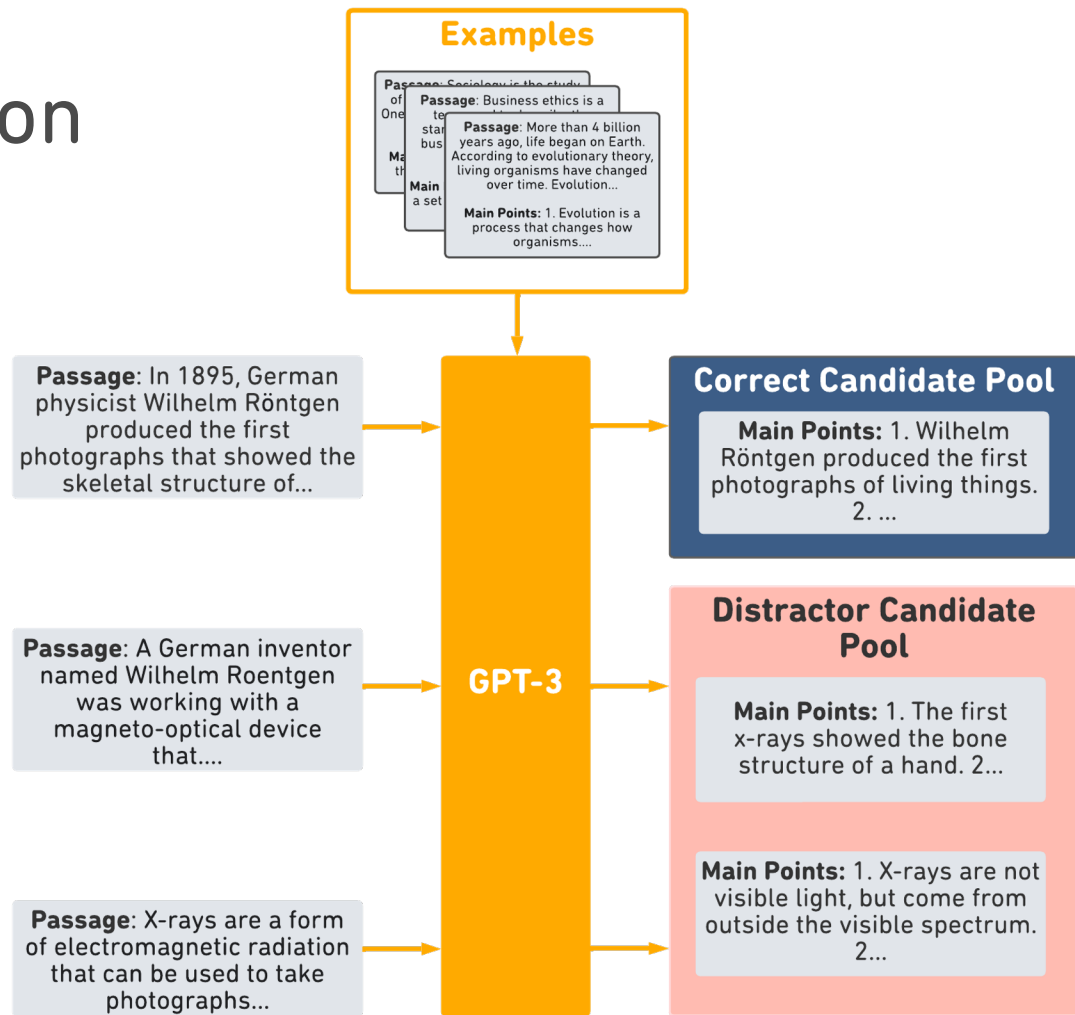
Goal:

Generate correct and incorrect options

Method:

Use of the alternative paragraphs as input prompts to generate answers, used as distractors

Generate alternative paragraphs on the same topic/title



Cloze item generation

- Use BERT (a different state-of-the-art LM) to sequentially propose most likely words at each word position
- Select words to hide based on
 - BERT analysis (was the most likely word the original word in the passage)
 - Lexical and syntactic analysis (part-of-speech, word frequency)
 - Linguistic context analysis (e.g., word repetition, dependency parsing)
- Select distractors based on
 - BERT likelihood
 - Lexical and syntactic analysis

Human review

Content Review at Scale: Overview

- Automated
- Item Quality
- Fairness and Bias

Item quality review



The purpose of item quality review (IQR) is to make sure that items target the skills and abilities in a manner that is consistent with the purpose of the item in order to reduce the likelihood of construct irrelevant variance. The main activities at this stage of review include reviewing and editing:

Fairness and bias review



The purpose of a fairness and bias-reduction (FAB) review is to ensure that all test takers have an equal opportunity to succeed. To that end, we want to eliminate potentially controversial content, materials, and subject matter that:

- may elicit negative reactions from marginalized group(s),
- contains stereotypical portrayals of groups of people, and
- would advantage or disadvantage test takers with certain backgrounds.

The DET should reflect a global perspective that does not unduly center one geopolitical or cultural body over another, e.g., avoid Eurocentricism or an implicitly United States perspective.

Content Review at Scale: Reviewer Portal

Current Ongoing Assignments

Name	Type	# of reviewers	Due Date	Review State	Update State
2022 Q3 Dictation Week 2.2	FAB	3	Tue, 30 Aug 2022 23:59:00 GMT	OPEN	
2022 Q3 Dictation Week 3	FAB	3	Tue, 06 Sep 2022 23:59:00 GMT	OPEN	
2022 Q3 C-Test Week 1	COPYEDIT	3	Tue, 30 Aug 2022 23:59:00 GMT	CLOSED	
2022 Q3 Listen-Repeat Week 1	COPYEDIT	2	Tue, 23 Aug 2022 23:59:00 GMT	CLOSED	
2022 Q3 Listen-Repeat FAB Week 1.1	FAB	3	Fri, 09 Sep 2022 23:59:00 GMT	OPEN	
2022 Q3 Listen-Repeat FAB Week 1.2	FAB	3	Tue, 13 Sep 2022 23:59:00 GMT	OPEN	
FAB Test	FAB	3	Sun, 31 Jul 2022 23:59:59 GMT	OPEN	
2022 Q3 Dictation Week 4	FAB	3	Tue, 13 Sep 2022 23:59:00 GMT	OPEN	

Menu

New
Review

Contractor
Home

Who reviews?

Item Reviewers

- Extensive experience in TESOL/Applied Linguistics, training in item writing
- Also trained to identify FAB issues

FAB Reviewers

- Diverse backgrounds - gender, country of origin, age
- Experience and interest in Diversity, Equity, Inclusion and Belonging (DEIB)

Reviewer management: Activities & Metrics in the Portal

Gather information about:

- Passage content labels
- Degree of content edits
- Selection/rejection of distractors, keys, comprehension questions
- Time taken to fix passages
- Decision agreement

Use results to improve the models to:

- Filter out bad/high effort generated passages
- Improve key and item distractor generation and selection
- Automatically fix certain content issues

Pilots

Pilot overview

Design

- 454 passages
- 5,246 items
- ~ 200,000 test takers
- ~425 responses / item



Pilot overview

Analysis

- Discrimination
- Item facility
- Response time
- Local item dependence



Item discrimination

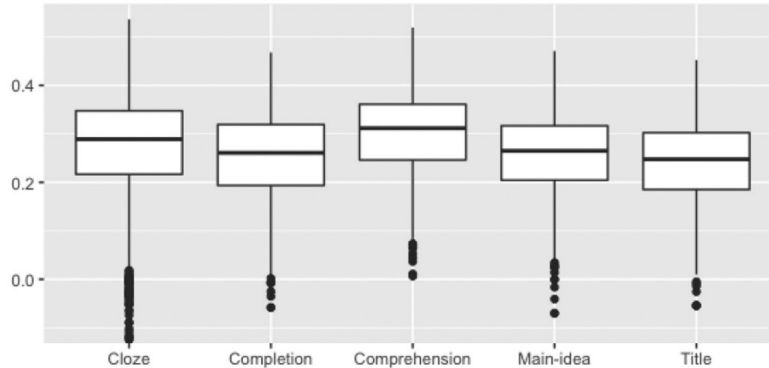


FIGURE 3
Item-total correlations with total practice test score.

Reasonably high

- Mean = 0.27
- 6% of items $r_{\text{item-total}} < 0.1$
- Distractor analysis can reduce this to less than 2%

Item facility

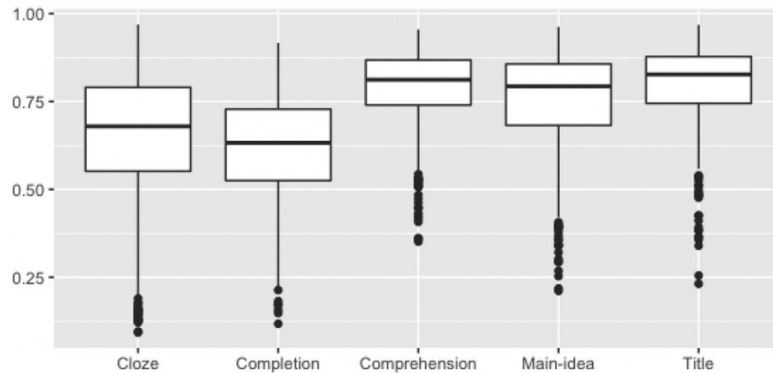


FIGURE 2
Mean score (easiness) distributions.

Wide facility distribution

- Mean = 0.70
- Earlier tasks (cloze & completion) less easy than later tasks

Response time

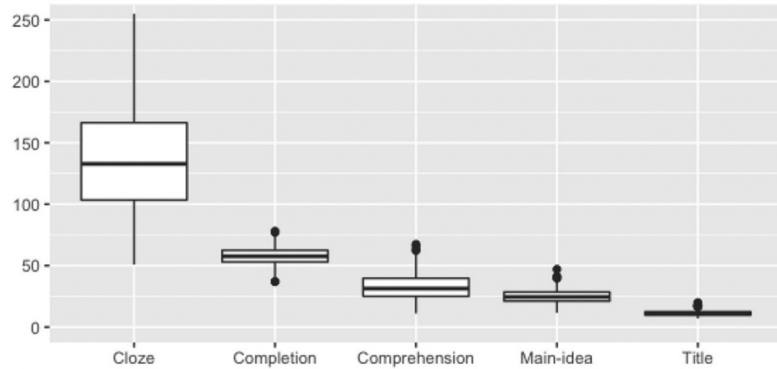


FIGURE 1
Median response time distributions (in seconds).

Response time ↓

Each subsequent item is completed more quickly than the previous item

~~Local item dependence~~

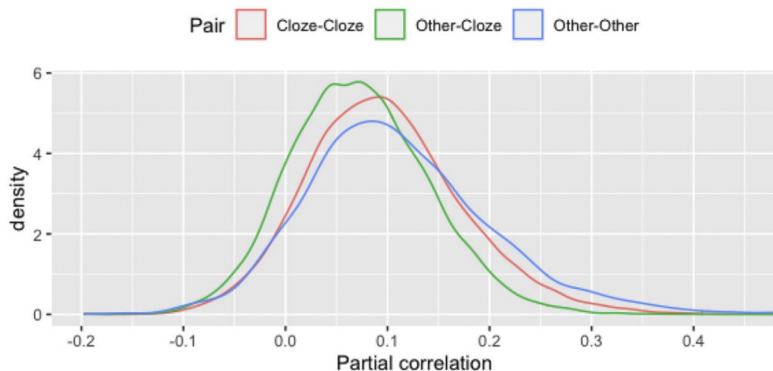


FIGURE 4
Partial correlations between item pairs.

Mean partial correlations

- Cloze-Cloze = 0.10
- Other-Other = 0.11
- Other-Cloze = 0.07

LID > 0.3

- 1% Cloze-Cloze pairs (> 9,437)
- 3% Other-Other pairs (> 4,540)

Wrap-up

Technology & task creation

Evaluation

- Interactivity and staged reveal contributes to decreasing response times
- This does not come at the expense of local item dependence
- Wide range of difficulty (important for adaptive administration)
- Linguistically similar to target domains

Result: Allows us to more efficiently scale our test creation processes

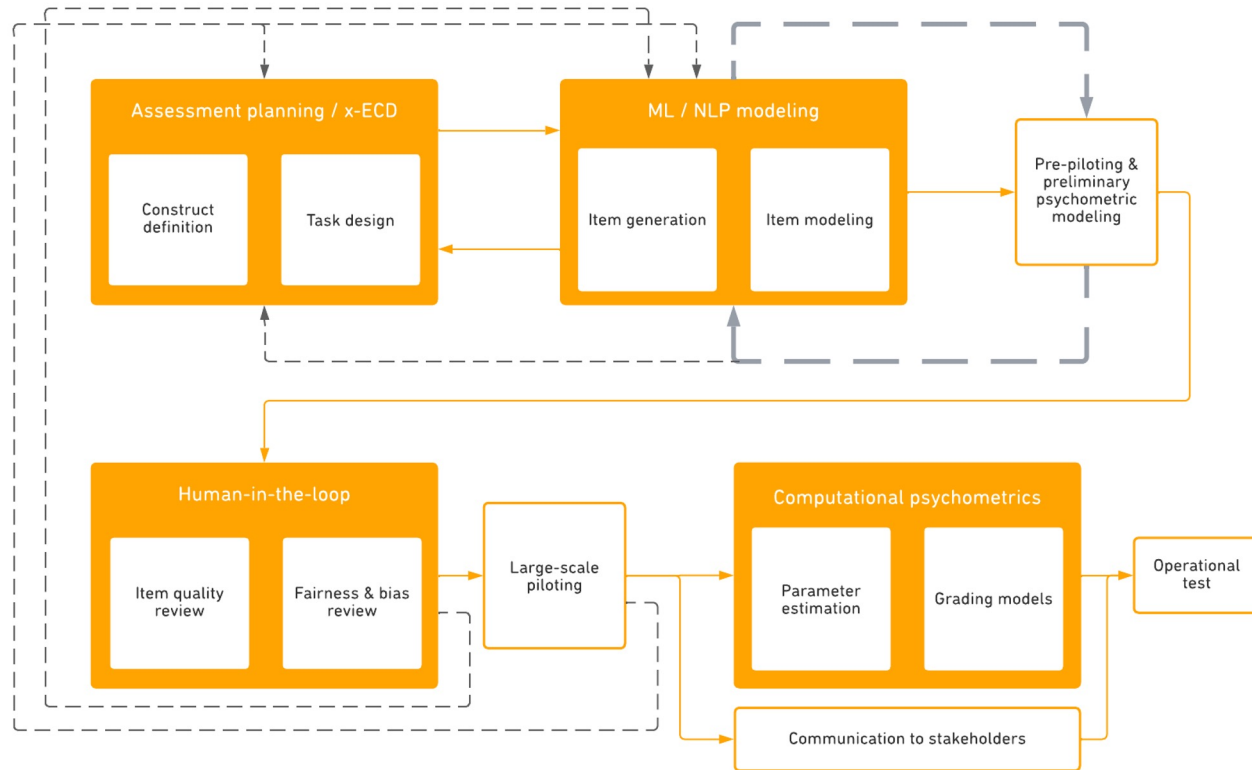
- Improve construct coverage through rapid prototyping of new items
 - Build a larger item bank to keep the test secure
 - Keep development costs low

Technology & task creation

Current & future applications

- Generate content for reading passages (and beyond)
- Generate test questions
- Maintain quality with human review
- Edit or rewrite content to meet a specific need
- Identify content that may not meet our testing guidelines

Digital-first test development process



Thanks!

englishtest.duolingo.com/research

geoff@duolingo.com