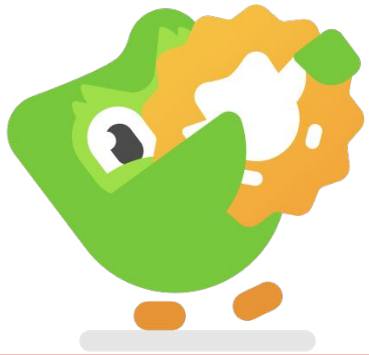
The magic of machine learning: How to develop tests in multiple languages simultaneously and at scale

Sarah Goodwin, Lauren Bilsky, Phoebe Mulcaire, and Burr Settles

Duolingo



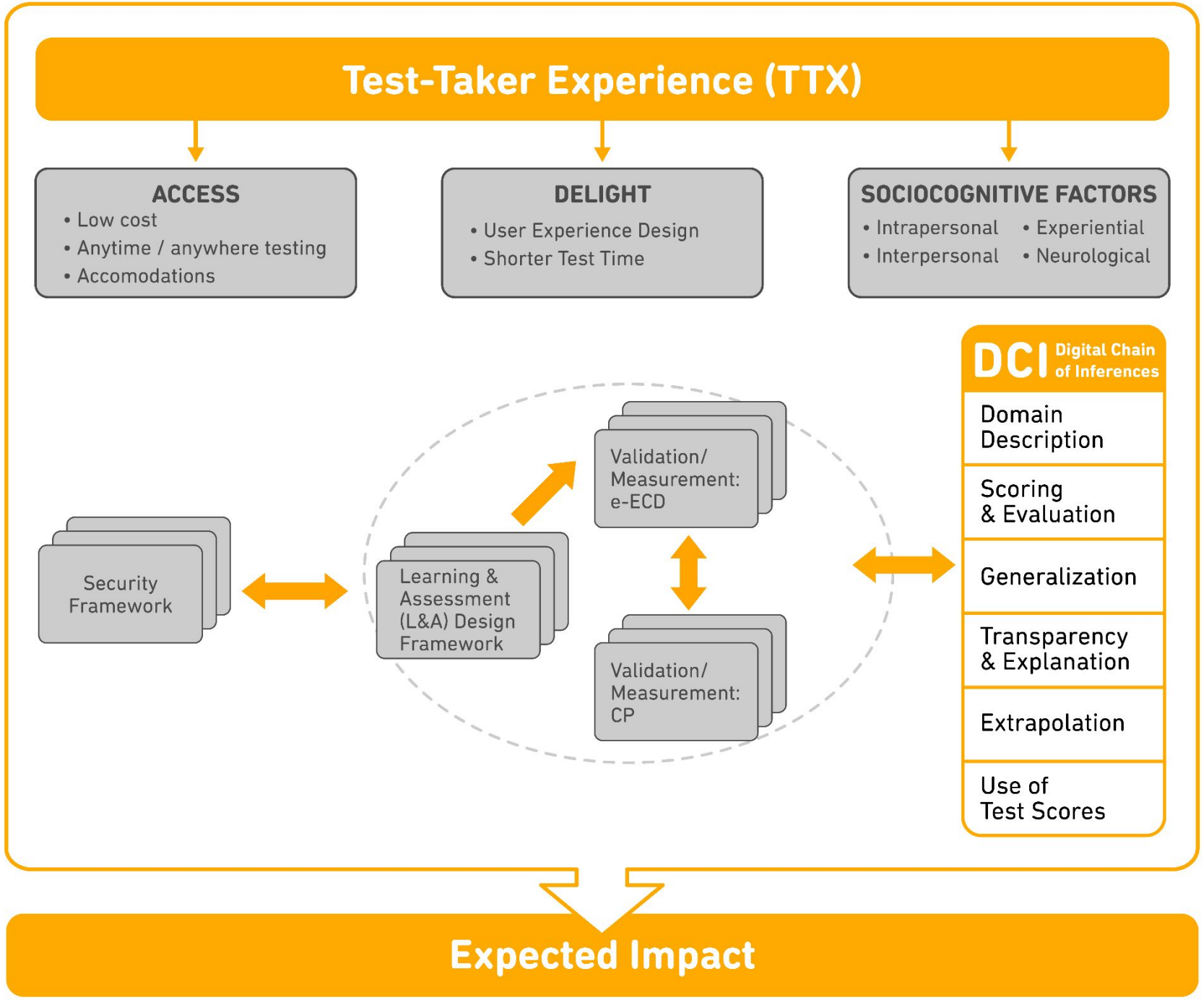
8th International Conference
MADRID
26-28 APR 2023
ALTE

In the project, I describe how we:

- Designed initial English and Spanish tests during two different company-wide "Hackathons". The English one is now what grew into the Duolingo English Test

Drew on those projects to, in 9 months:

- Scale existing infrastructure: multilingual applications of item generation and difficulty estimation
- Design prototype FR and ES pilot tests (listening and reading tasks)
- Pilot with two groups of examinees
- Analyze results to compare item responses and machine-learning difficulty estimates


Test-Taker Experience (TTX)
ACCESS
• Low cost
• Anytime / anywhere testing
• Accomodations
DELIGHT
• User Experience Design
• Shorter Test Time
SOCIOCOGNITIVE FACTORS
• Intrapersonal
• Experiential
• Interpersonal
• Neurological
Security Framework
Learning & Assessment (L&A) Design Framework
Validation/ Measurement: e-ECD
Validation/ Measurement: CP
DCI Digital Chain of Inferences
Domain Description
Scoring & Evaluation
Generalization
Transparency & Explanation
Extrapolation
Use of Test Scores
Expected Impact

Prototype test traits

- Scored computer-adaptive test (CAT) for receptive skills of **listening and reading**: yes-no audio-vocabulary, yes-no printed text-vocab, dictation, and c-test
  - Listening: Word recognition, vocabulary knowledge
  - Reading: Word recognition, vocabulary, meaning from discourse/context
- **Short**: full exam takes less than 30 minutes
- **Low stakes**: no proctoring
- ML modeling to **predict text- and word-level difficulty even before item administration** (applying Multilingual BERT [roberta-xlm-base] and methods from Settles et al., 2020; McCarthy et al., 2021)
  - BERT = Bidirectional Encoder Representations from Transformers, transformer-based machine learning

0:55

Select the real French words in this list

sécurités	tamaille	rapporatit	encourager	rebond	flanquer
flacain	vasside	rédaction	caractiques	guetter	renfosse
baldrer	probuls	sourcil	kiroug	cadonnée	étranchir

NEXT

1:27

Select the real Spanish words in this list

WORD 1	WORD 2	WORD 3
WORD 4	WORD 5	WORD 6
WORD 7	WORD 8	WORD 9

NEXT

Yes/no vocabulary curation

Real words:

- Learning app courses (only A-B levels), Cervantes resources for ES (A,B,C-level), OpenSubtitles, and GoogleBooks
- ML predictions and CEFRlopod
- Filtering
  - English words, names, places, profanity
  - Part-of-speech (POS) prediction
    - Not always accurate

Nonwords:

- Generated for each language separately
- Plausible phonology and orthography
- Trained character recurrent neural network on large set of words
- Filtering
  - Real words, profanity, too unlikely
  - Homophones (using model from Duolingo Speech Lab)

Word-level difficulty estimation

- Classification on expert-labeled CEFR-labeled real words
- Needed to predict apparent difficulty for nonwords - so used features that are defined for fake words
  - no word frequency, POS
  - ngram language model, length, num syllables
- Same model for text and audio yes-no vocabulary

1:32

Type the statement that you hear

Your response

Number of replays left: 2

NEXT

2:57

Type the missing letters to complete the text below

La Antigua Roma es el nombre de una civilización en Italia. C o m como u pequeña c o m u agrícola e el s i VIII a.C. S convirtió en una c i u y tomó e nombre d Roma de s fundador Rómulo. C r e hasta convertirse en el m a imperio d mundo a n t . Comenzó como un reino, luego se convirtió en una república, luego en un imperio.

NEXT

Sentence and passage difficulty

- Dictation sentence-level difficulty – short decontextualized sentences – so classification based on internal tools that estimate CEFR level
- C-test text difficulty is at the passage level
  - With responses, can add word-level information
- Multilingual BERT (roberta-xlm-base)

Pilot 1 of 2

Recruited Mechanical Turk participants:

- Took language acquisition/use survey; did not use French or Spanish extensively in formative years
- Items too difficult, so restricted item pool to easiest items

Pilot 2 of 2

Recruited learners from Duolingo language learning app:

- Needed to have progressed far enough in 10 units (v1): Units 3 to 10
- 1097 ES and 599 FR unique target language examinees
- Examinees each took 6 items (max total time 9 minutes):
  - First item on every test session was an “anchor” text vocab item
  - Then one c-test, two dictation, one textvocab, and one audiovocab, which could appear in any order

Learner pilot findings

- Unique items with 30+ or more responses: 156 Spanish; 87 French
- Item-level scores ranged from **0.062** to **0.976**
- **Learners' highest unit reached in learning app** (3 to 10) and **session scores** correlated at Pearson's *r* = **0.325** (*p*=0.000)

Learner pilot findings, cont'd.

- **Average item scores** and **machine-learning item difficulties** were moderately correlated:
  - *r* = **-0.521** for Spanish, *r* = **-0.404** for French
  - ES c-test correls. not as strong as correls. for other 3 item types (though comparable to English ctest)
  - FR textvocab and dictation correls. stronger than for other 2 item types

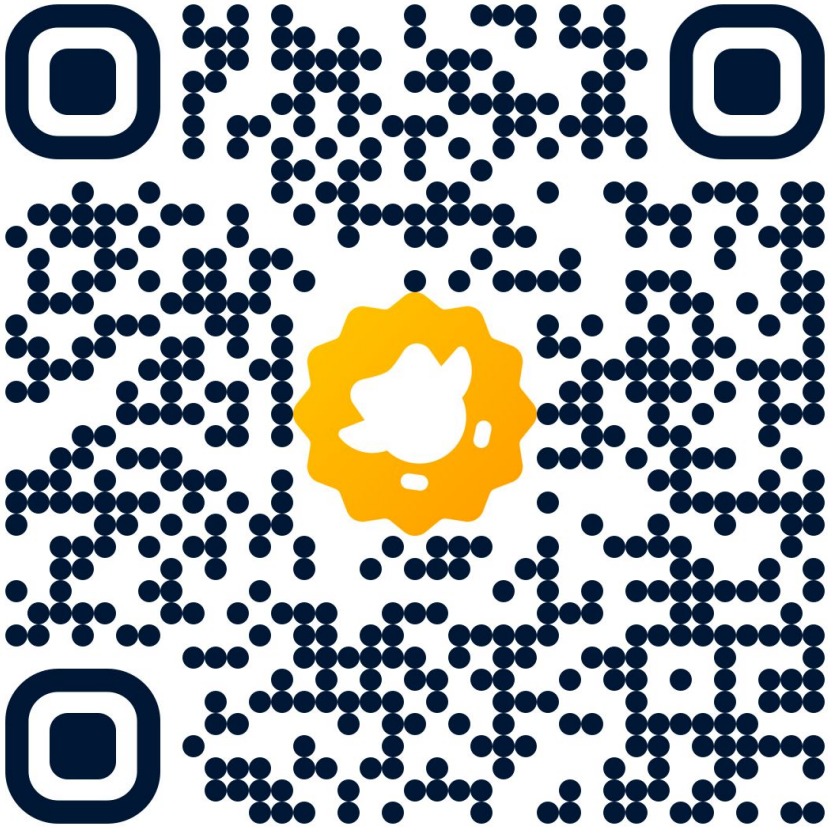
What we did: Produced proof-of-concept multilingual modeling for tests in other languages

- Adapted BERT using multilingual models to bootstrap difficulty estimates
  - BERT = Bidirectional Encoder Representations from Transformers, transformer-based machine learning
- Crafted item rules and generated items for 4 CAT item types
- Deployed platform for people to take a complete test and receive an overall score (10 to 160 scale)

Implications for test development

- low-cost, accessible options
- harnessing AI/ML expertise - interdisciplinarity
- future scaling to less commonly taught languages

¡Gracias! Merci !



Email SarahG@duolingo.com for framework paper

References

American Council on the Teaching of Foreign Languages [ACTFL]; The National Standards Collaborative Board (2017). *NCSSFL-ACTFL Can-Do Statements*. https://www.actfl.org/resources/ncssfl-actfl-can-do-statements

Blanco, C. (2021, December 6). *2021 Duolingo Global Language Report*. Duolingo. https://blog.duolingo.com/2021-duolingo-language-report

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/abs/1907.11692

McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021, November). Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 883-899).

Settles, B., T. LaFlair, G., & Hagiwara, M. (2020). Machine learning-driven language assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247-263.