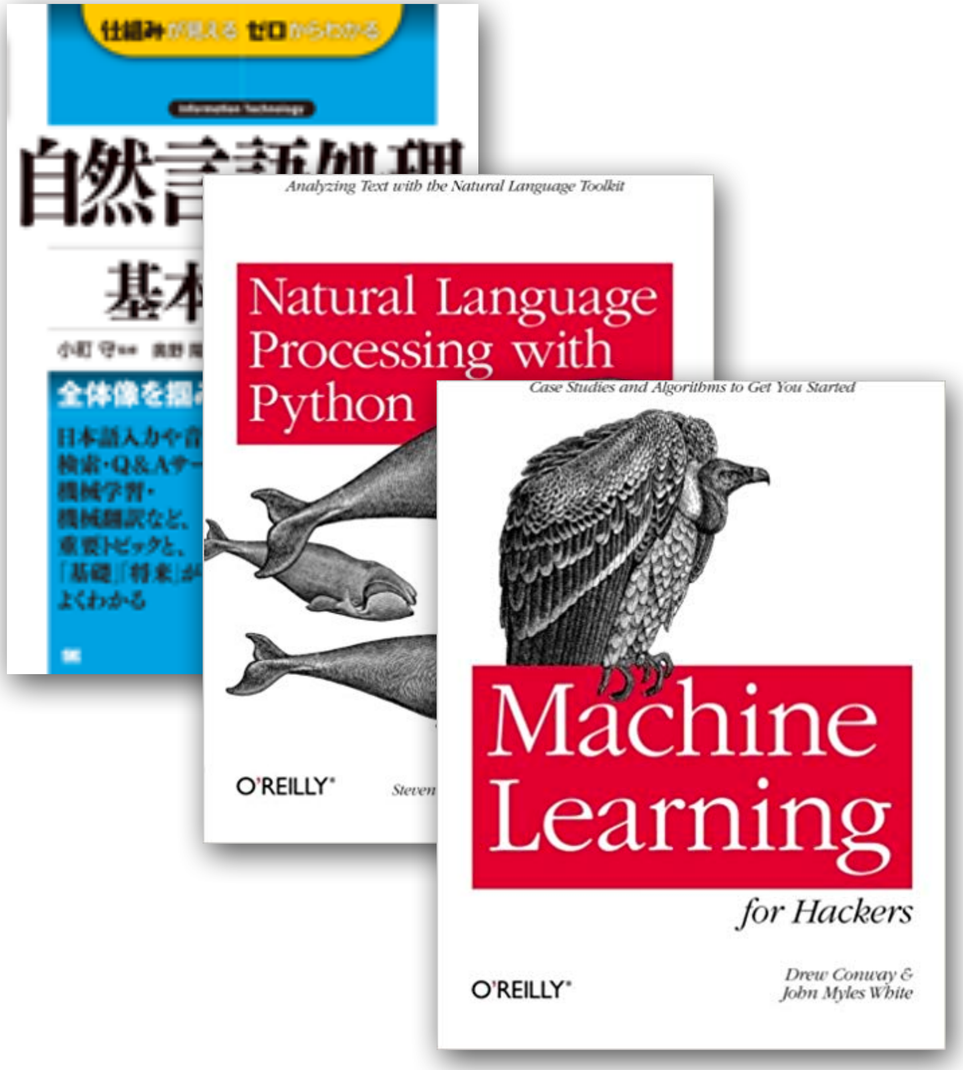
duolingo english test

Masato Hagiwara

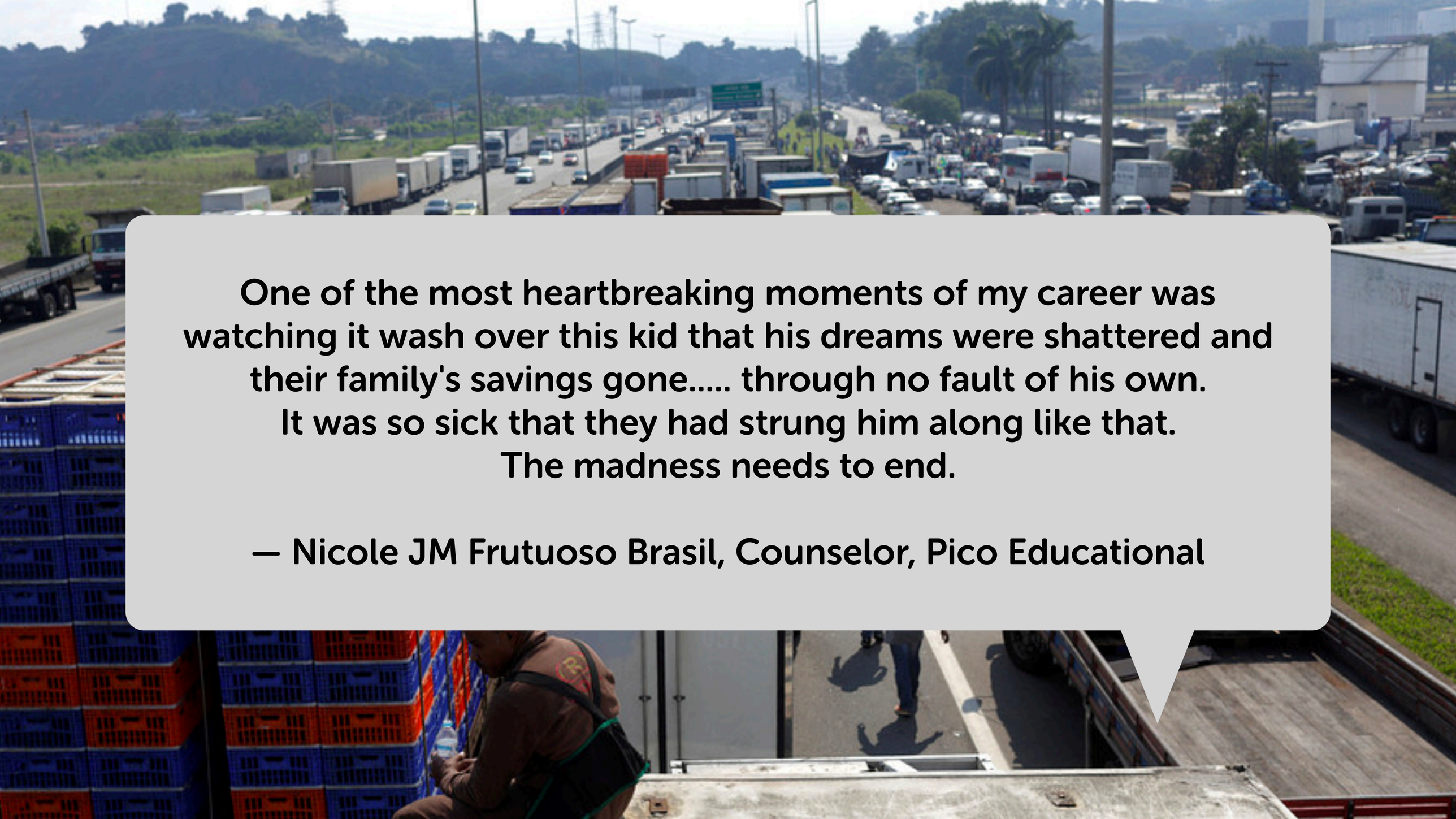
Senior Machine Learning Engineer / Researcher

- PhD in **AI/ML** with focus in natural language processing
- Joined Duolingo in **2015**
- Published three books on ML/NLP



Luis

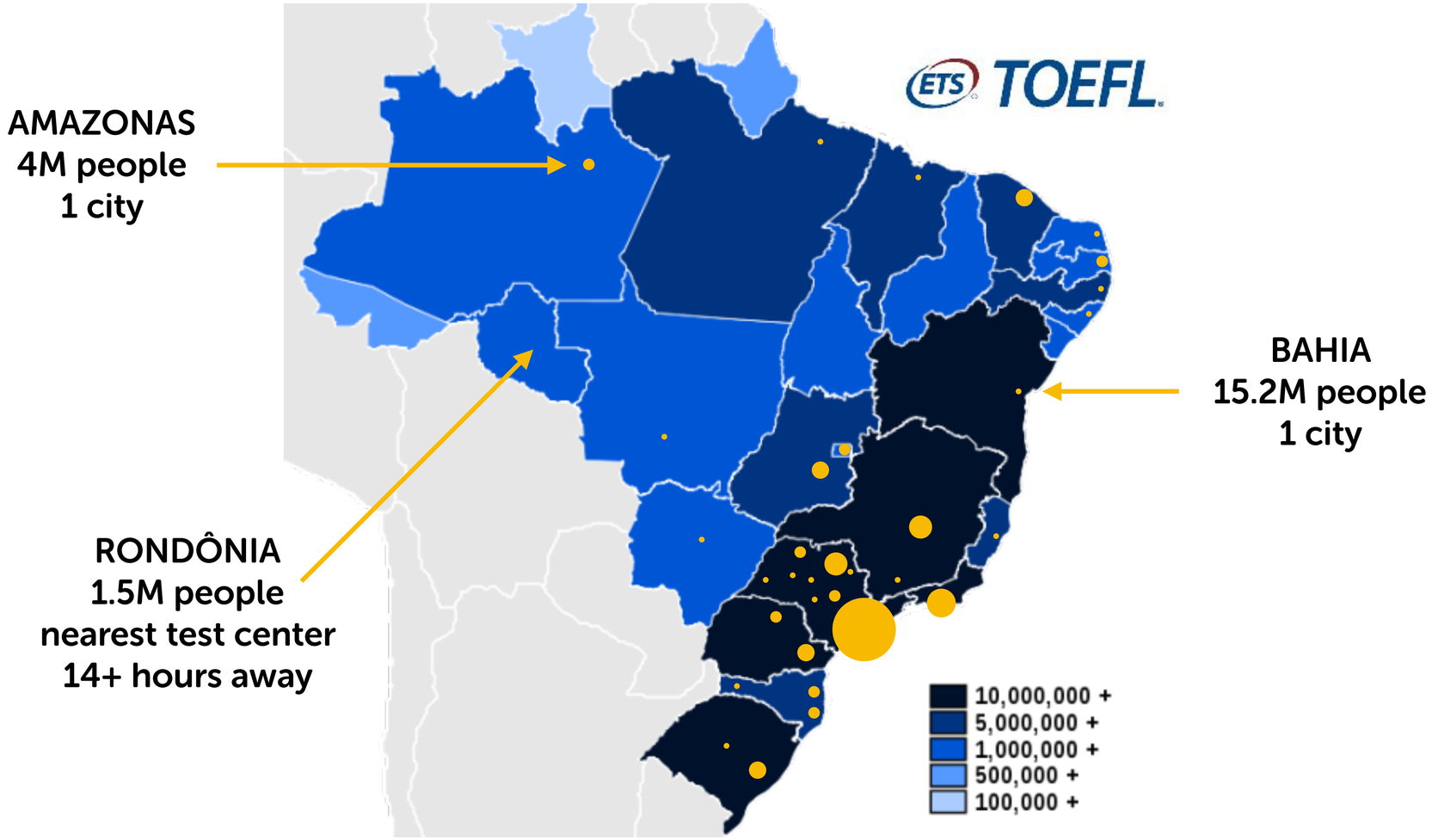
Grew up in **Guatemala**



One of the most heartbreaking moments of my career was watching it wash over this kid that his dreams were shattered and their family's savings gone..... through no fault of his own.
It was so sick that they had strung him along like that.
The madness needs to end.

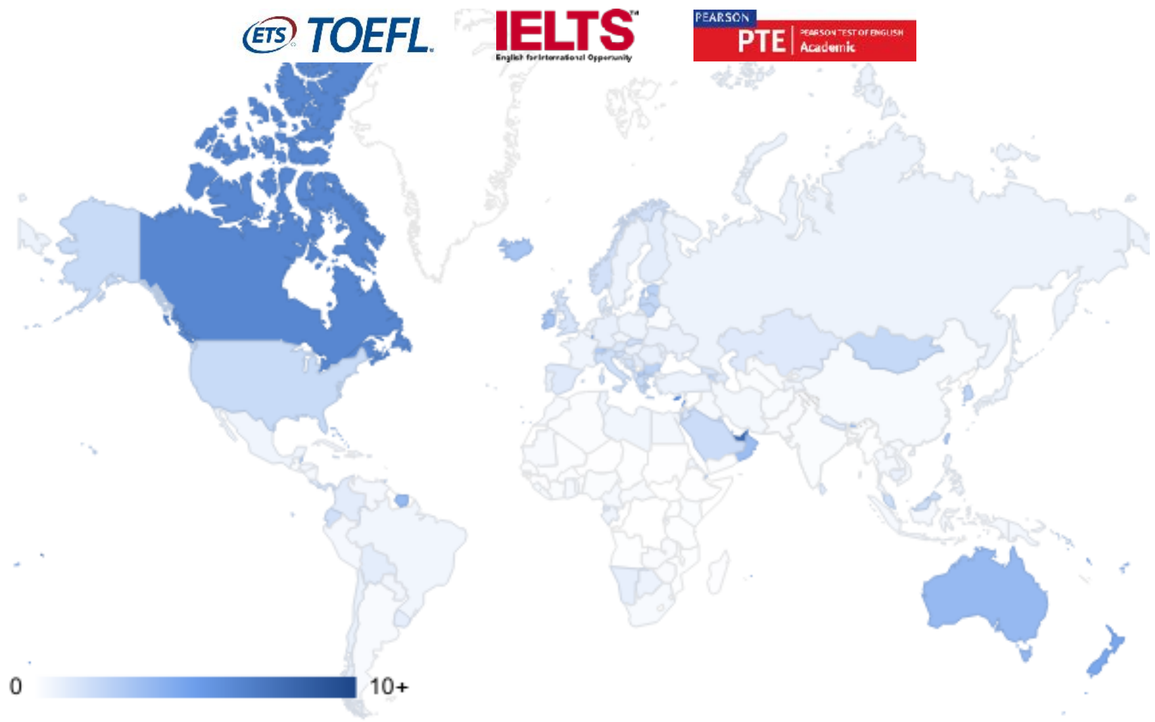
— Nicole JM Frutuoso Brasil, Counselor, Pico Educational

Test Centers in Brazil (July 2018)

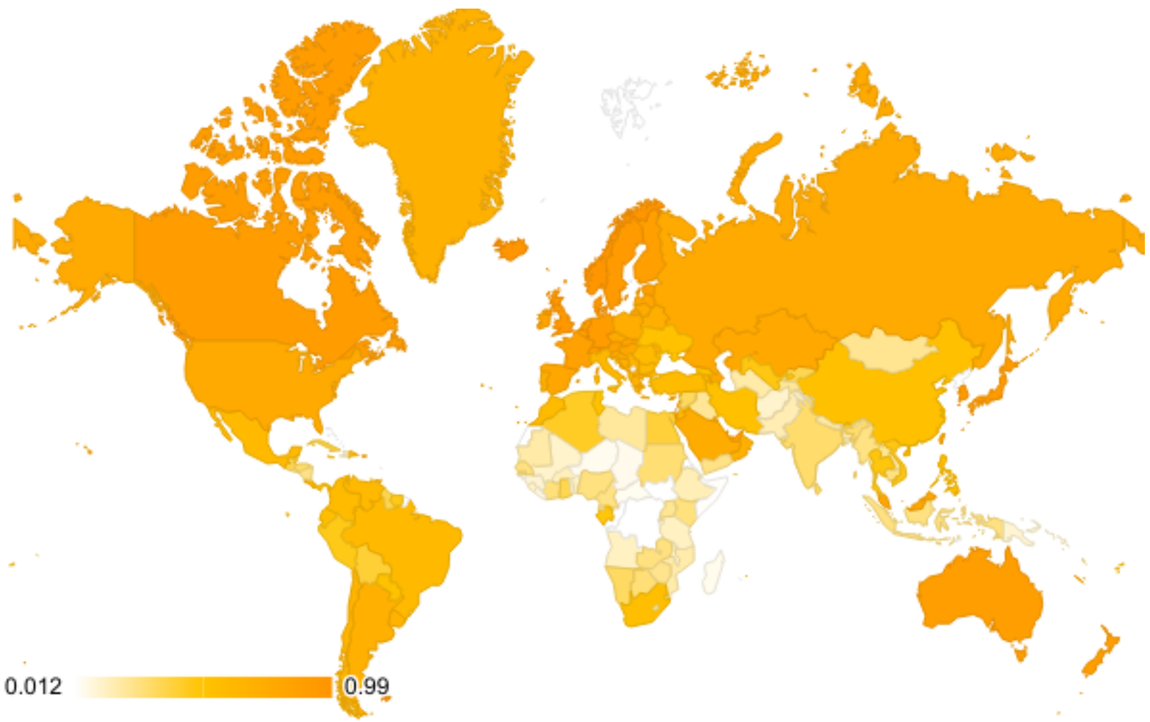


Test Centers vs Internet Access

1 Test Center per 1,926,260 people



1 Internet user per 2.2 people

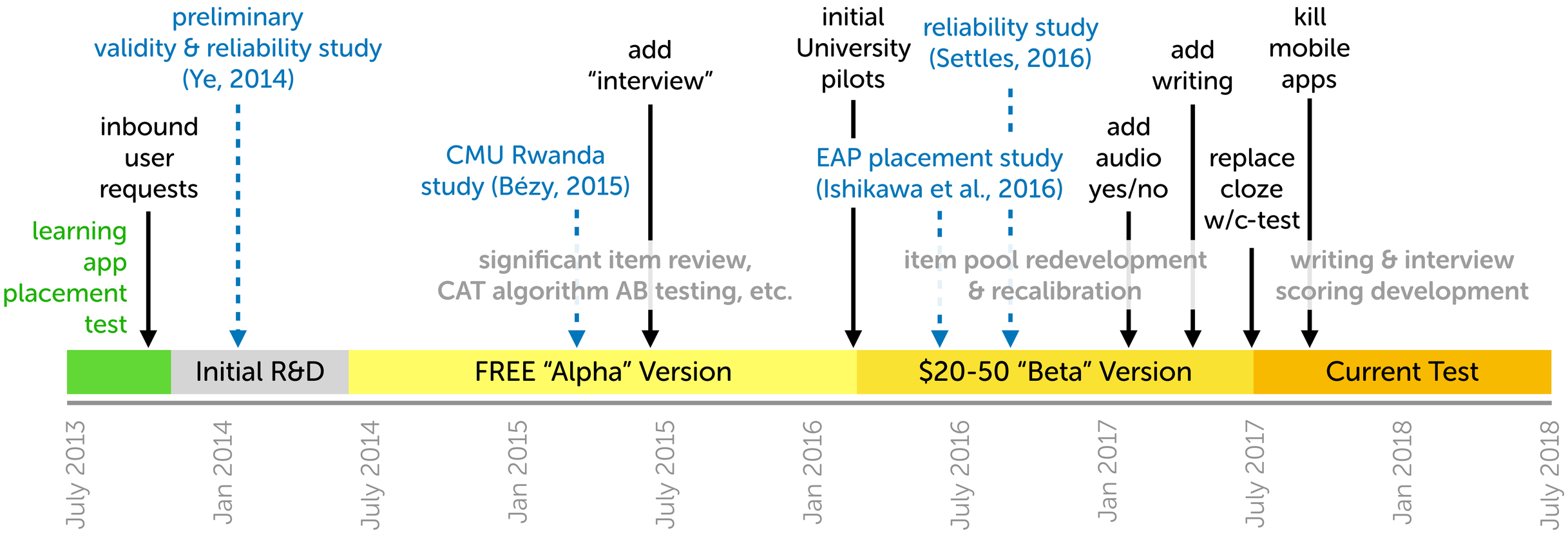


Sources:
https://www.ets.org/bin/getprogram.cgi?urlSource=toefl
https://www.ielts.org/book-a-test/find-a-test-location
https://pearsonpte.com/the-test/test-centers-and-fees/
International Telecommunication Union (ITU), 2016

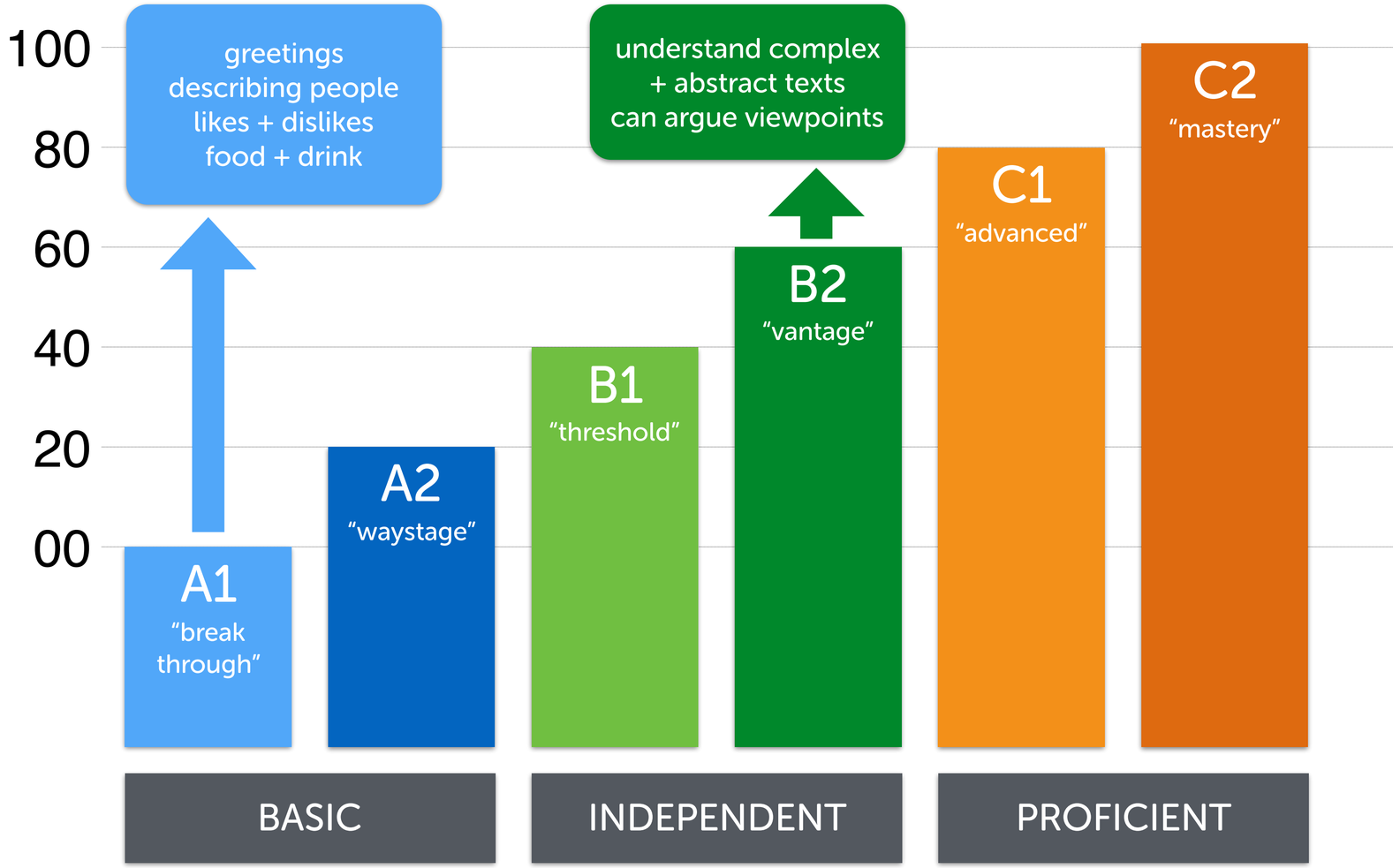
Can we build an accurate, affordable online English test to overcome these access barriers?

A Timeline of the DET

https://englishtest.duolingo.com/resources

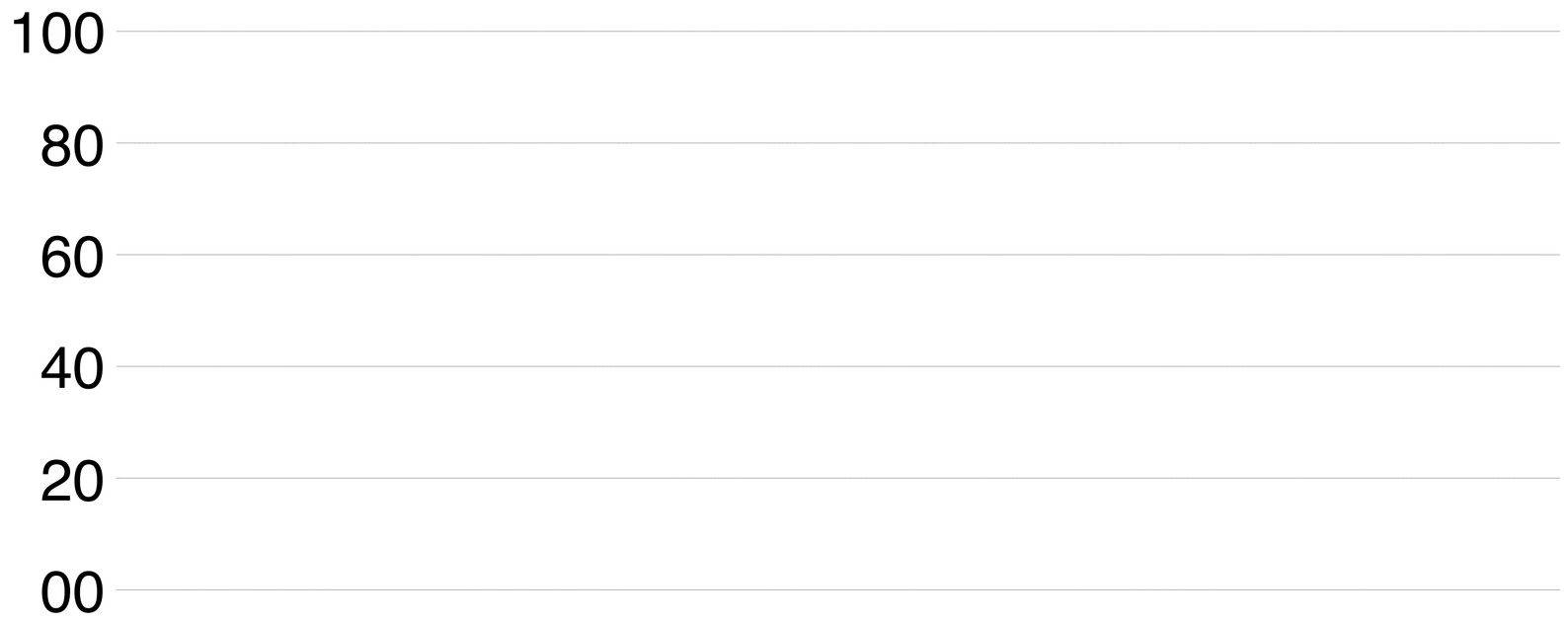


DET Construction



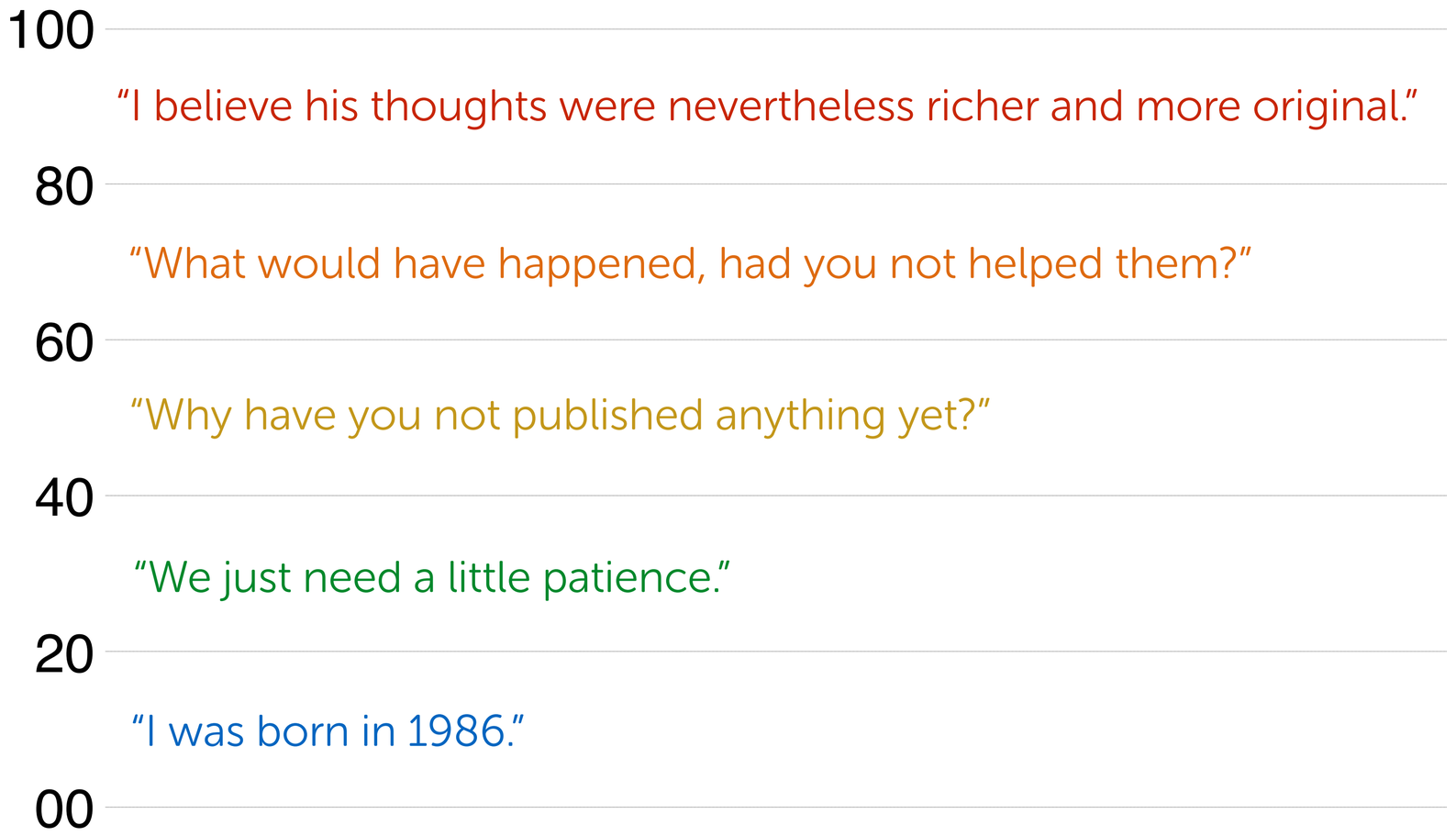
- Common European Framework of Reference

DET Design & Construction



- Common European Framework of Reference
- Duolingo scale

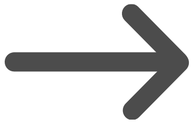
DET Design & Construction

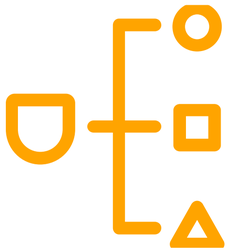


- Common European Framework of Reference
- Duolingo scale
- level estimation via ML/NLP

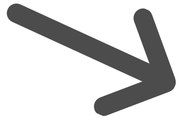
DET Design & Construction

Expert-annotated
CEFR texts & wordlists



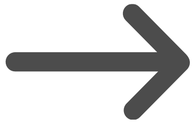


AI-based
CEFR difficulty estimation



Large pool of
CEFR-aligned
items for
CAT administration

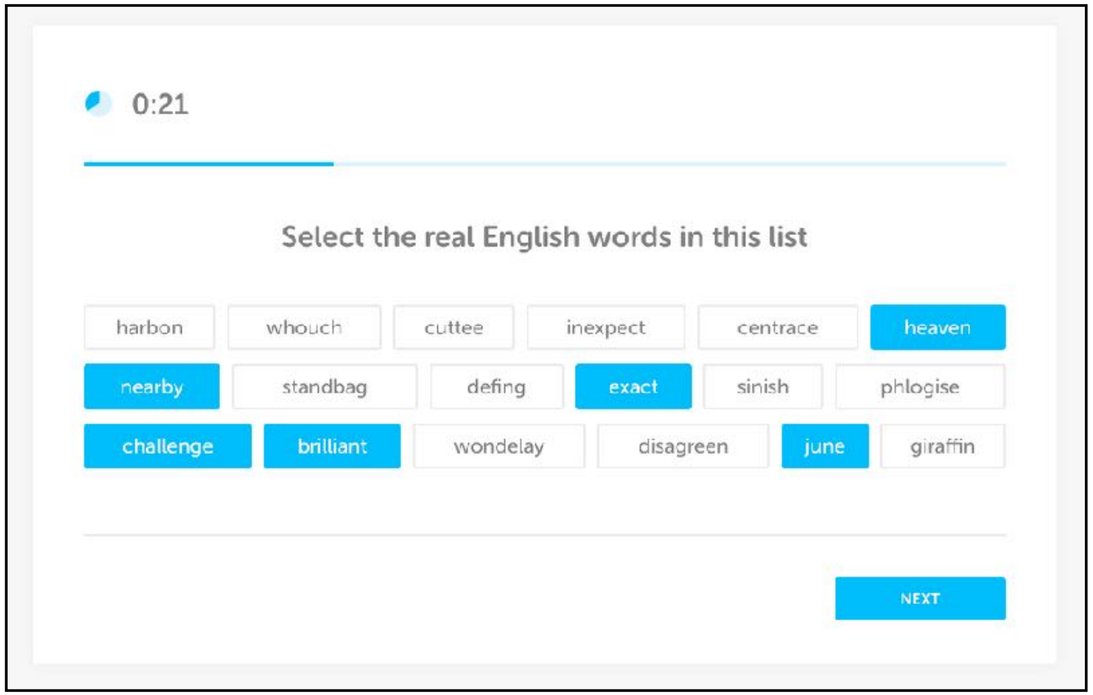
Natural English
language sources



Item-generation
techniques

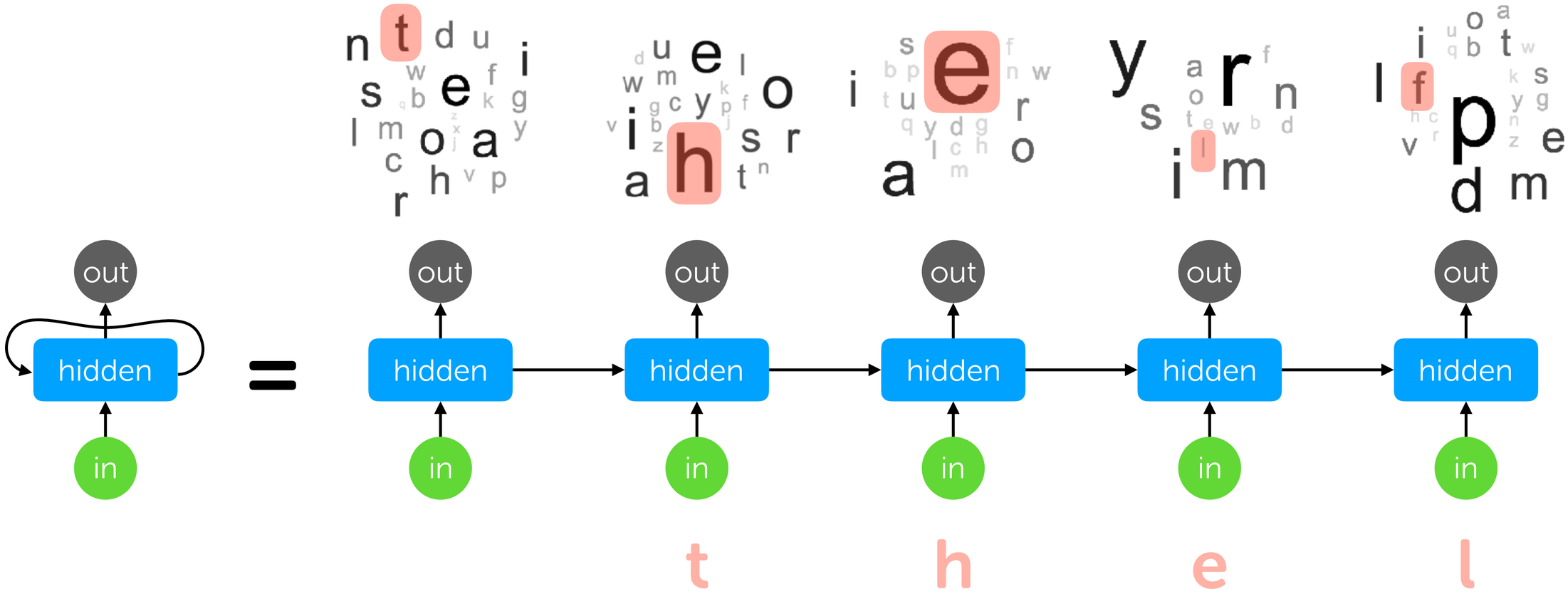
Deep Dive: “Yes/No” Vocab

- distinguish English **words** from English-like **pseudowords** (Zimmerman et al., 1977)
- **written variant** significantly predicts reading, writing, & listening skills (Milton, 2010; Staehr, 2008)
- **audio variant** significantly predicts listening & speaking skills (Milton, 2010)

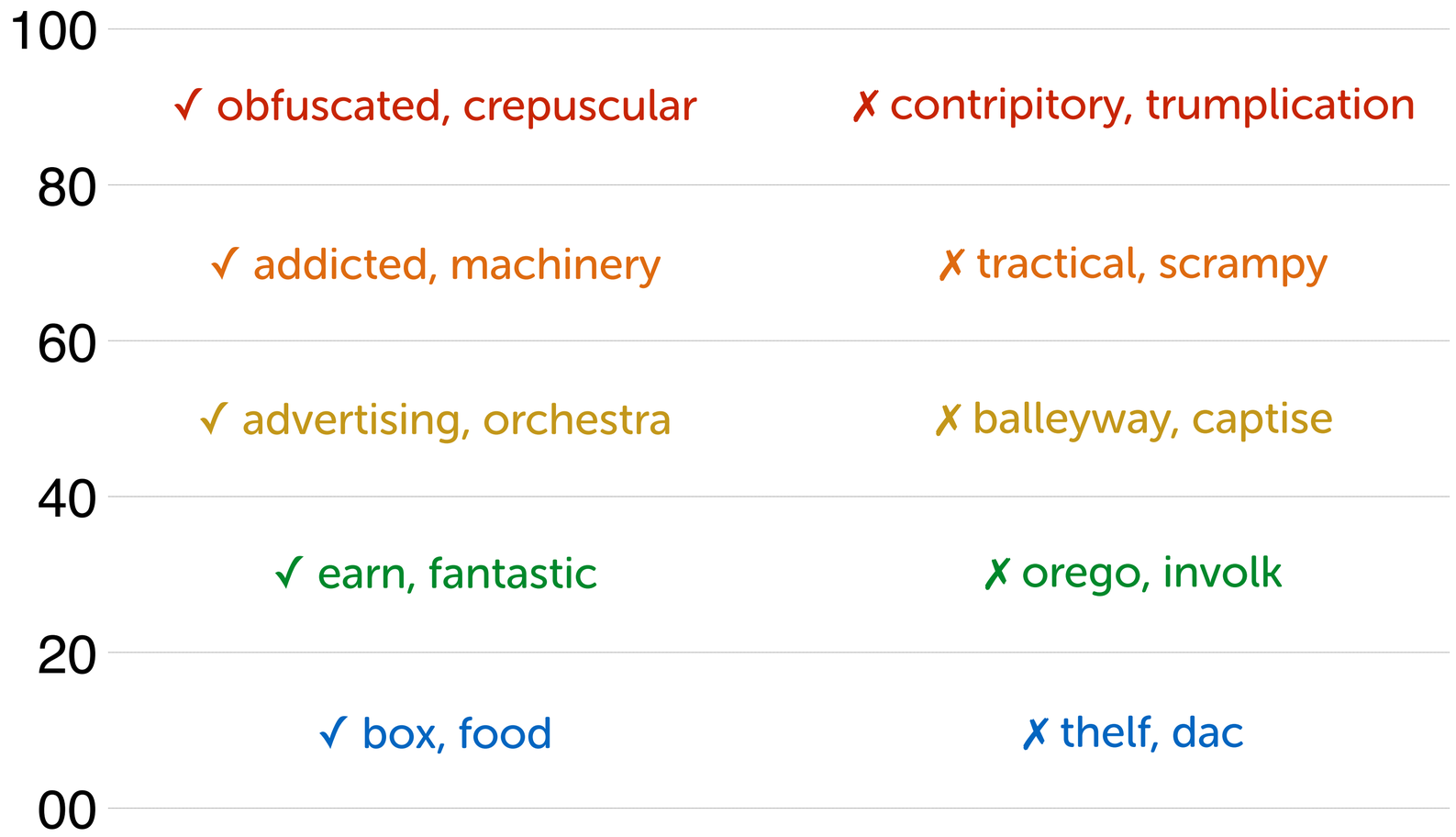


Generating Pseudowords

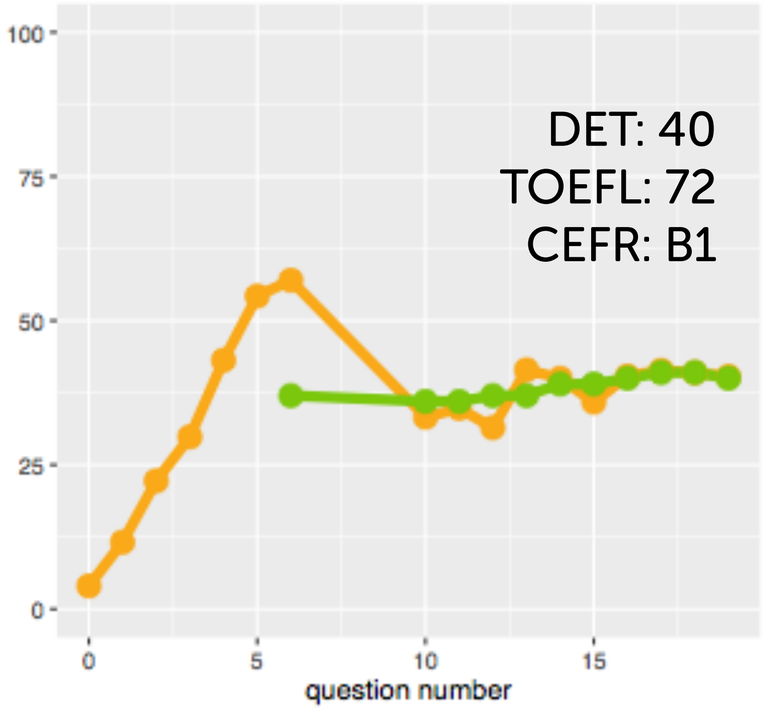
- we use **recurrent neural networks** (LSTM RNNs)

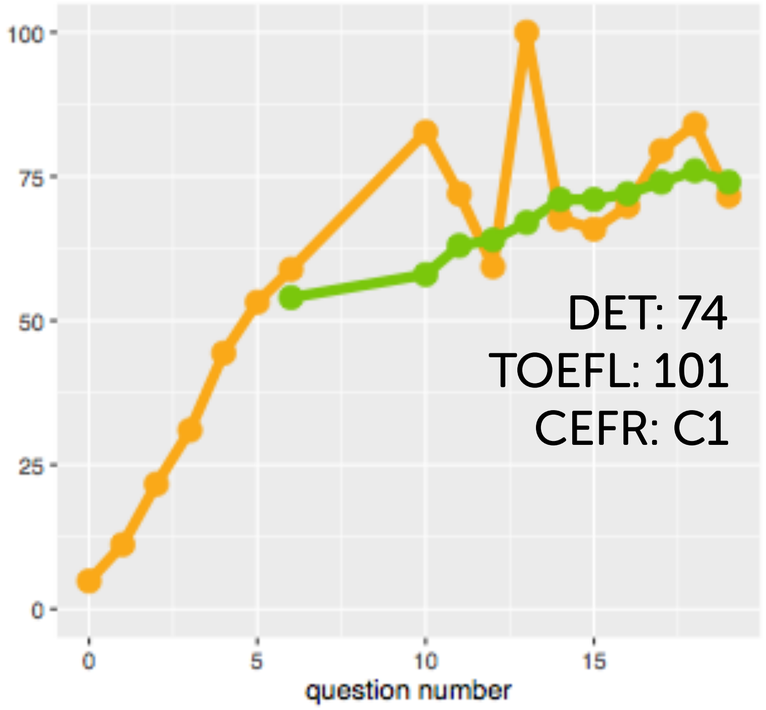


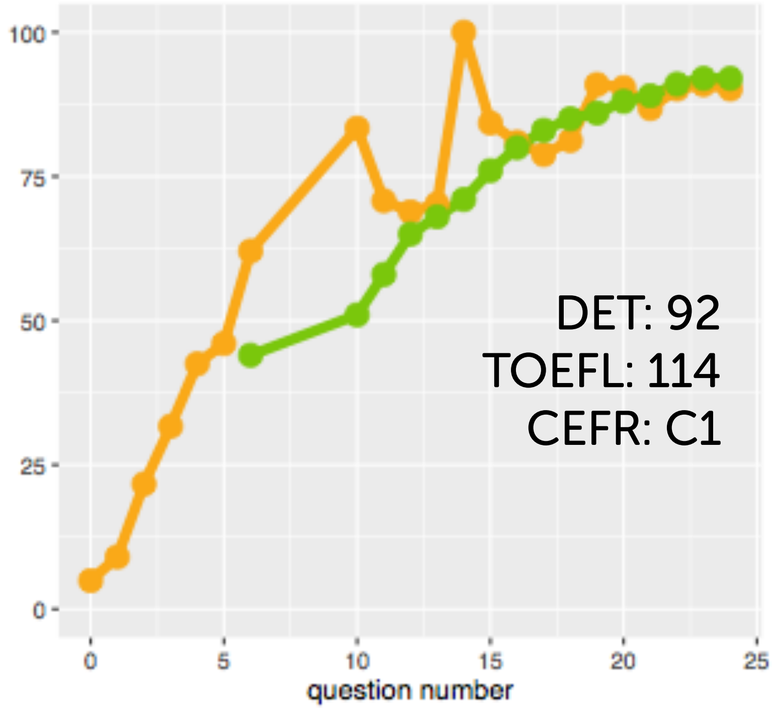
Projecting the Score Level



Computer-Adaptive Testing (CAT)

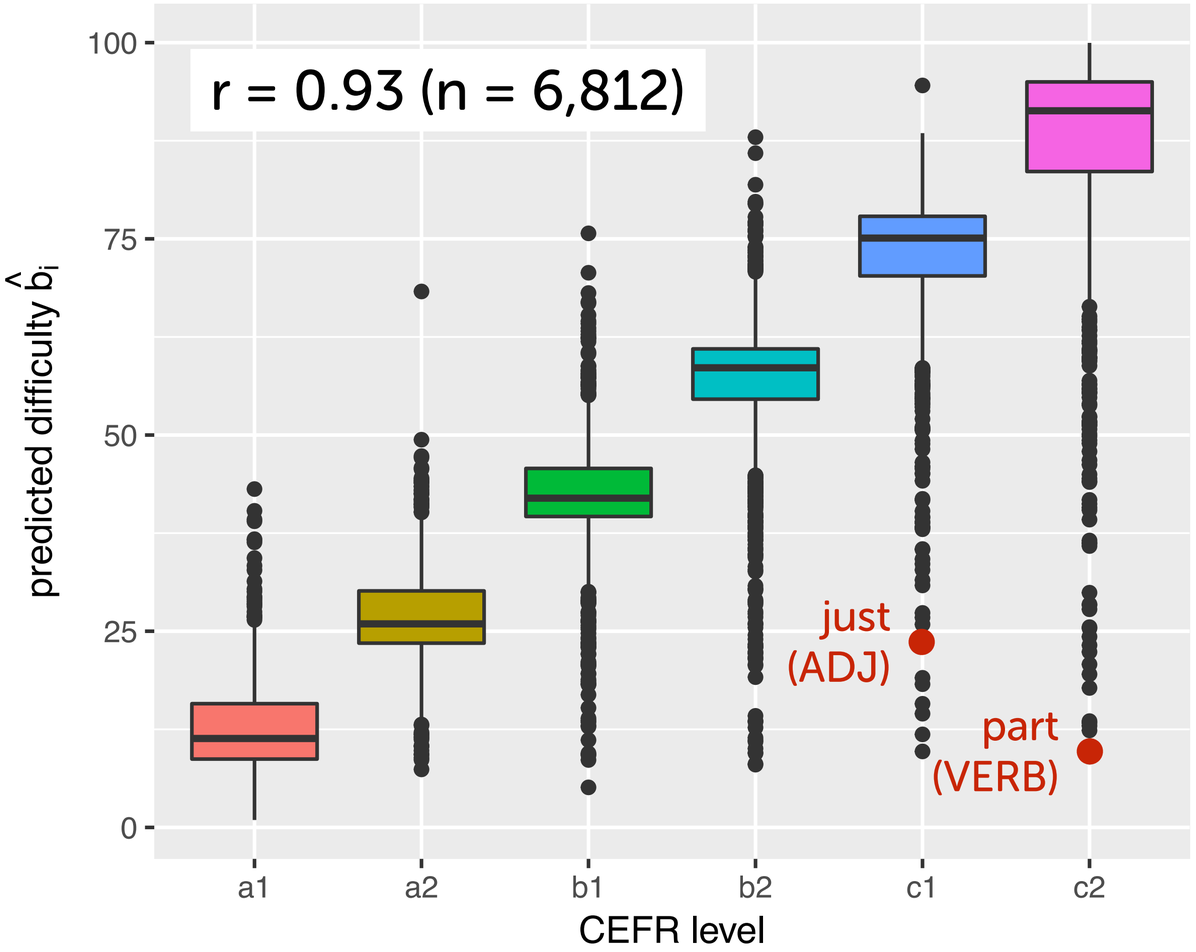






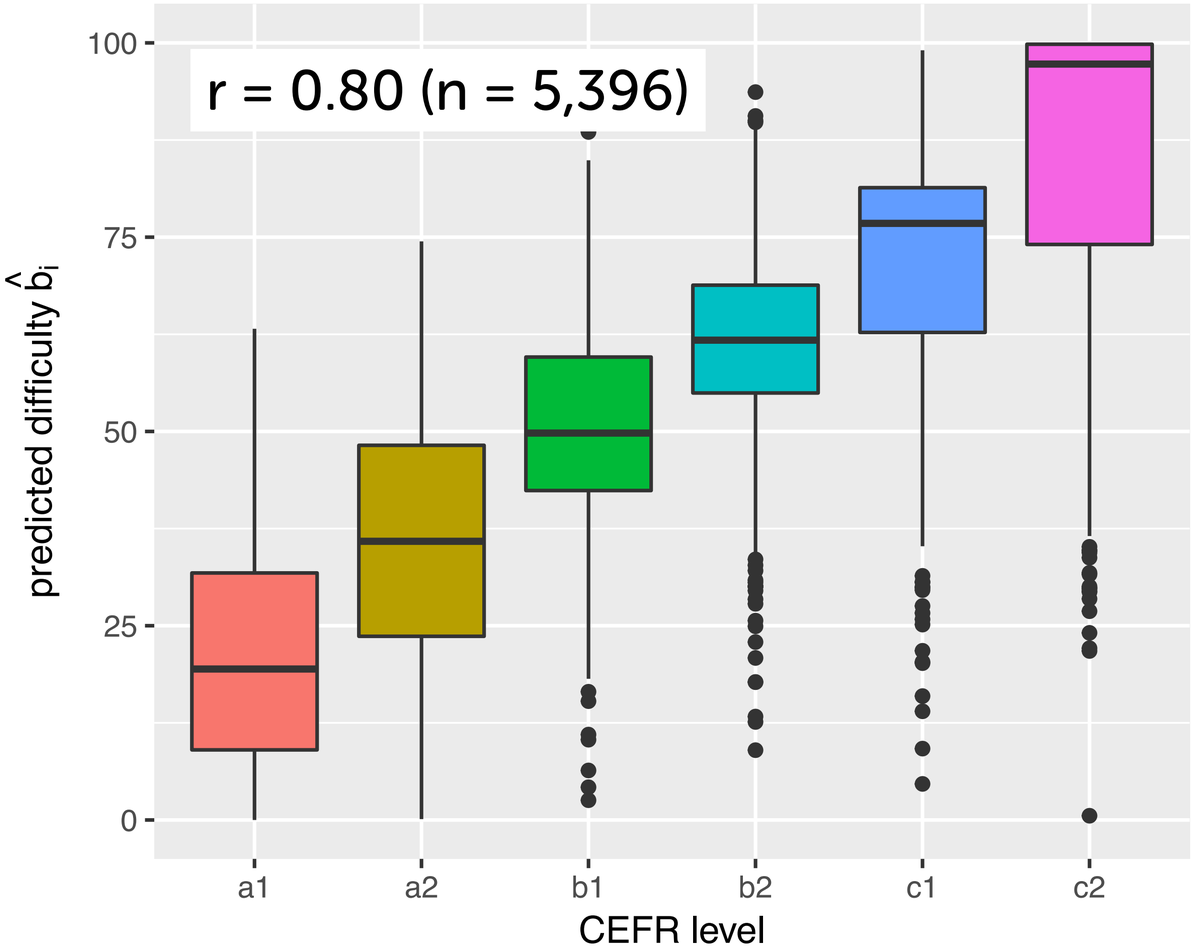
Validity Evidence

Relationship with CEFR (Yes/No Vocabulary)



(5-fold cross-validation; wordlists compiled from Capel, 2010/2012; North et al., 2011; other sources)

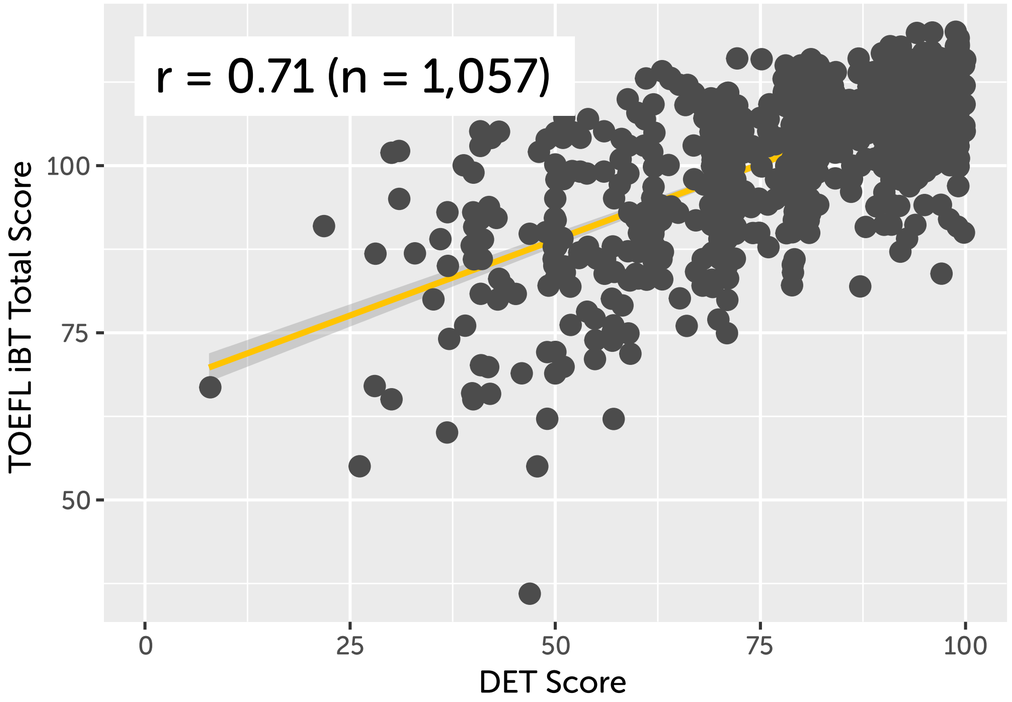
Relationship with CEFR (Passage)



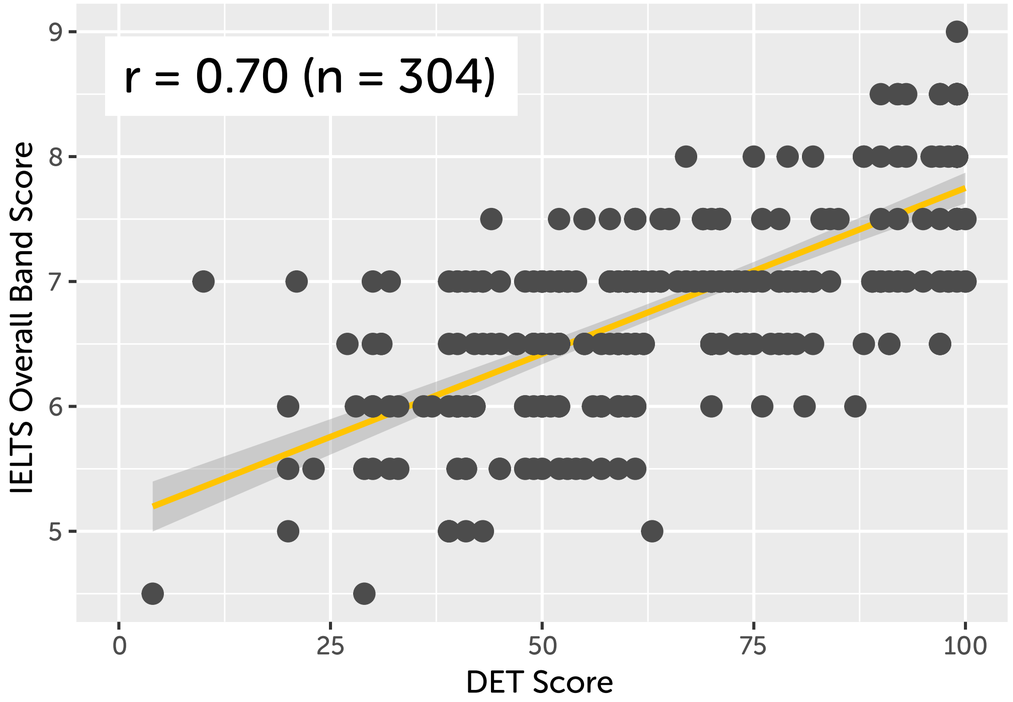
(5-fold cross-validation)

Relationship with Other Tests

DET↔TOEFL iBT



DET↔IELTS



TOEFL iBT↔IELTS

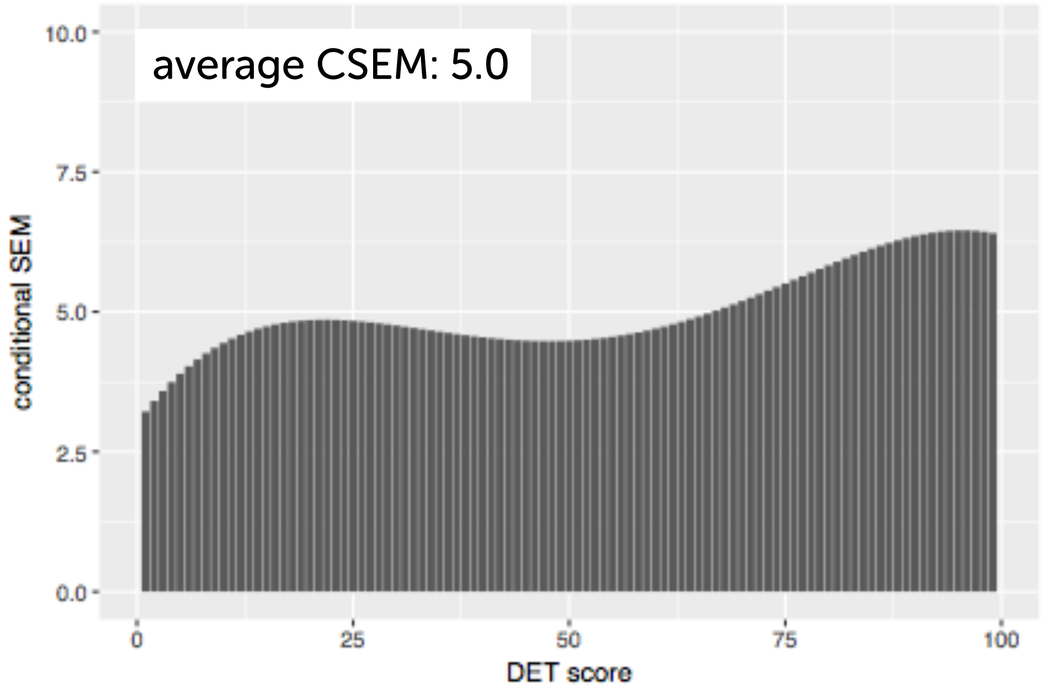
r = 0.73 (n = 1,153)

(ETS, 2010)

Reliability

internal consistency (split-half): 0.95

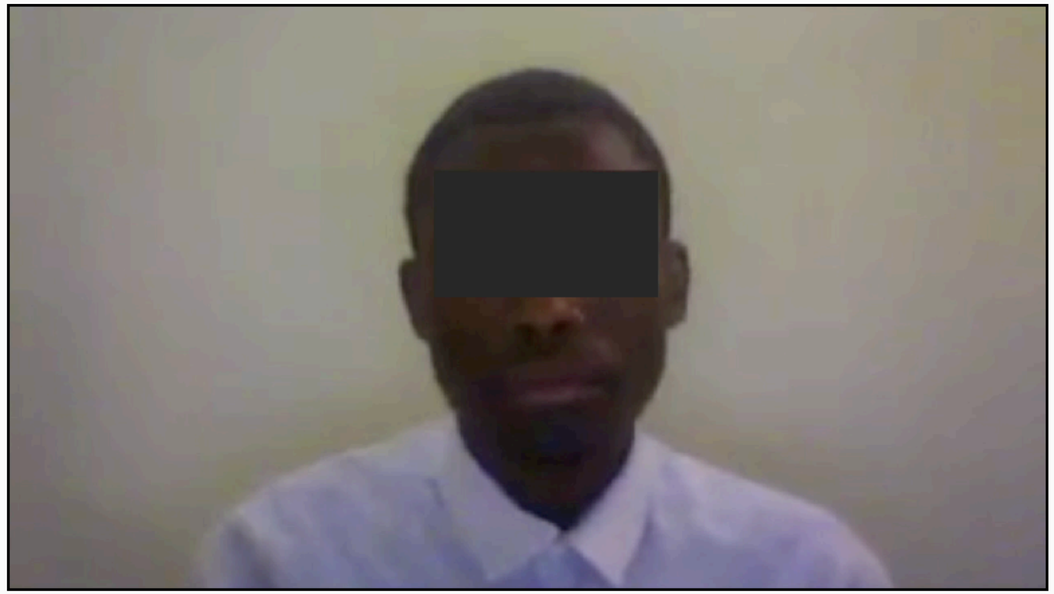
SEM-based: 0.96



test-retest: 0.85

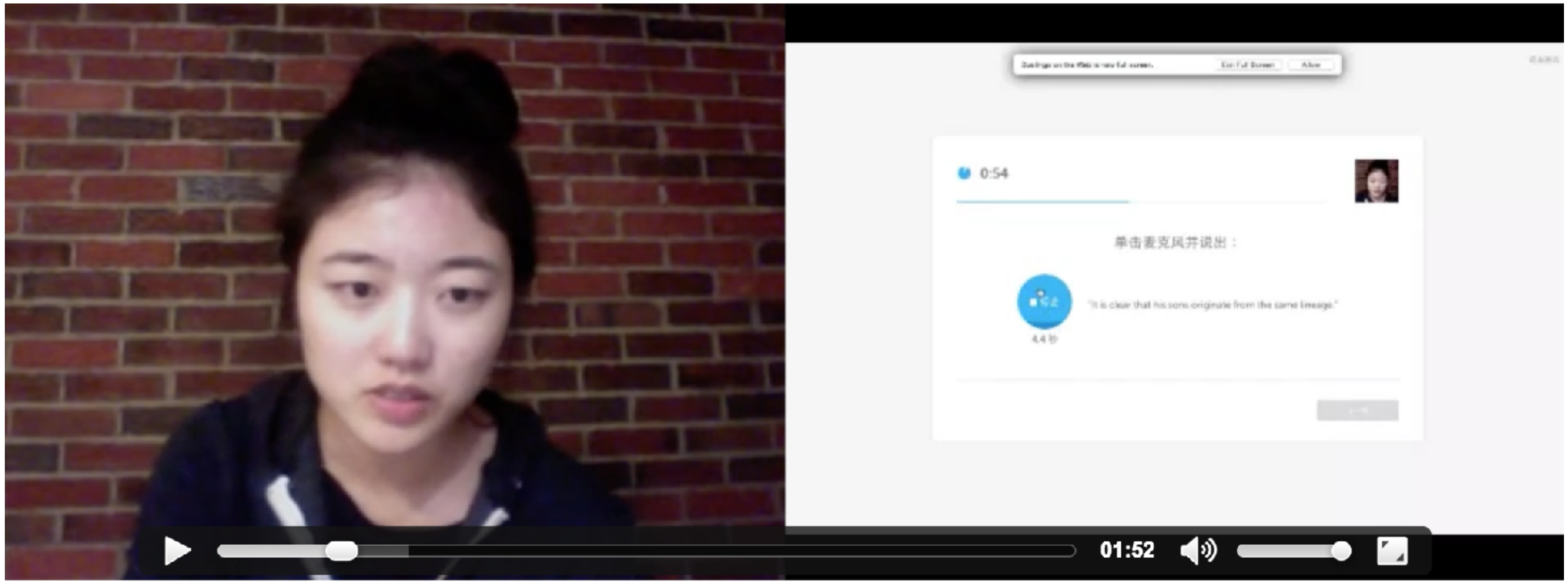
“Robert” from Zimbabwe

case study from US University: was **unable** to take TOEFL/IELTS
& SAT verbal was **below cutoff**, but otherwise very strong applicant



“The **DET definitely strengthened** what was probably the most concerning piece of his app.”

Security



item/testlet exposure rate	0.01%	(~1 in 1,000 tests)
exam overlap rate	0.81%	(<1 in 100 items shared)
examinee overlap rate	0.89%	(<1 in 100 items shared)

duolingo english test

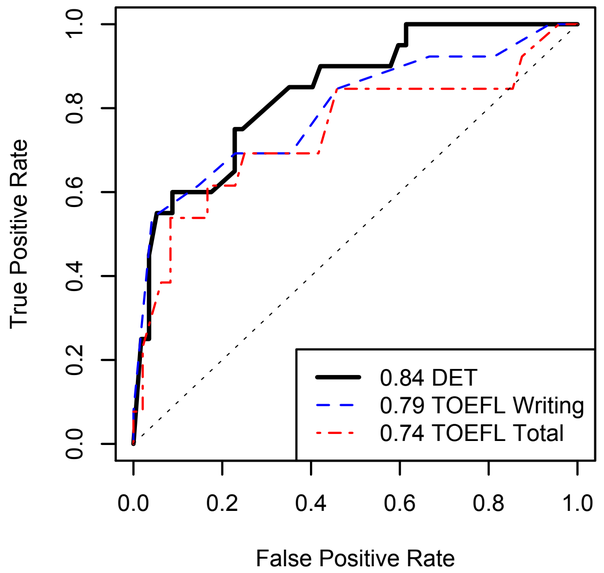
Thank You! Questions?

Relationship with EAP Faculty Placement

(Ishikawa et al., 2016)

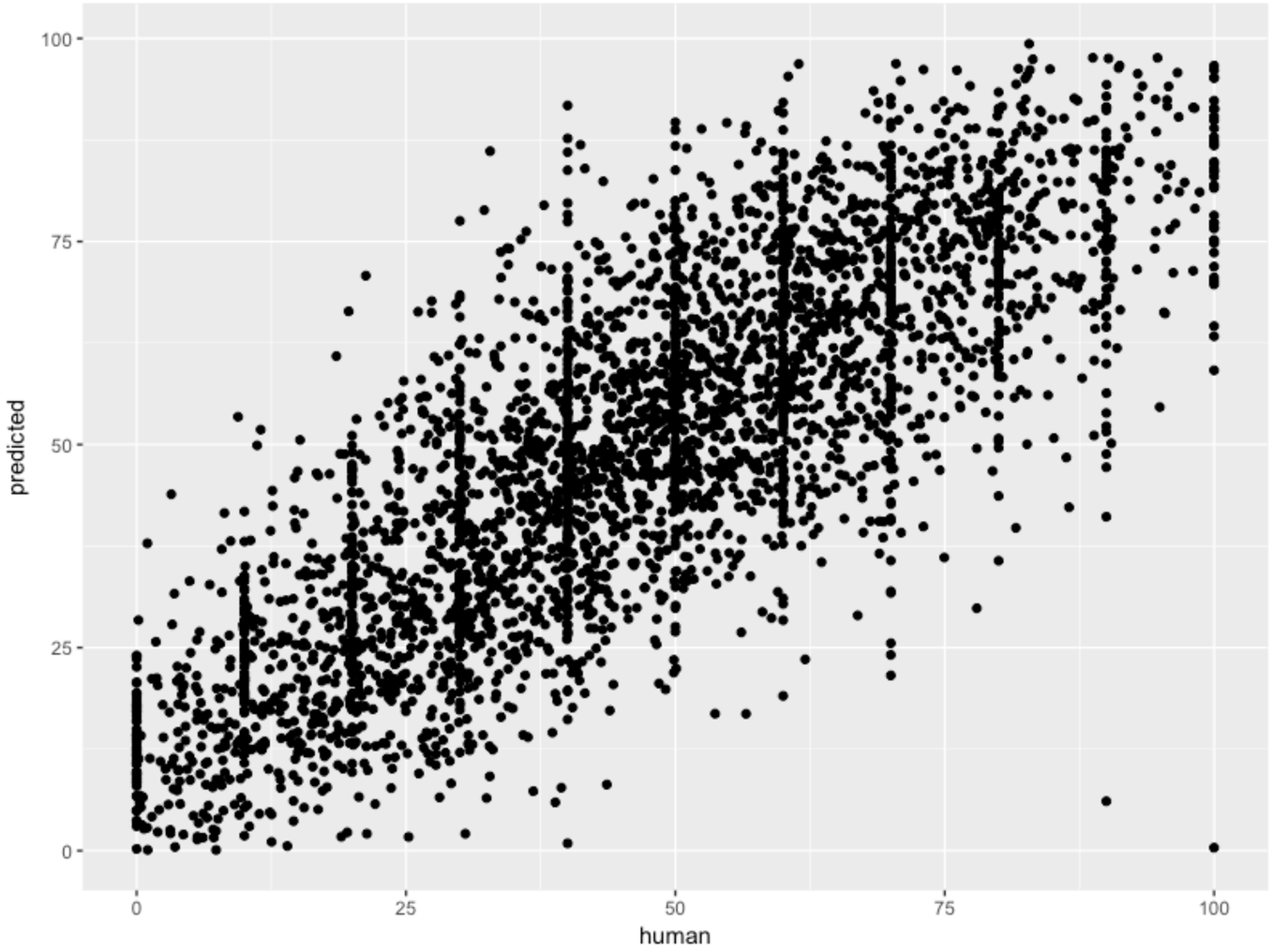
	2014 Cohort ($n = 31$)				2015 Cohort ($n = 46$)			
Faculty Assessment	DET	TOEFL	Writing	Speaking	DET	TOEFL	Writing	Speaking
Written	**0.69*****	0.12	0.47*	—	**0.58*****	0.36*	0.53***	—
Oral Comprehensibility	**0.40***	0.27	—	0.37	**0.52*****	0.26	—	0.40*
Oral Fluency	**0.45***	-0.09	—	0.13	**0.39****	0.13	—	0.36*
Oral Pronunciation	**0.45***	0.13	—	0.33	**0.42****	0.10	—	0.36*

Note: $^{***}p < 0.001$ $^{**}p < 0.01$ $^{*}p < 0.05$



ROC curves for EAP support class placement decisions

Automated Interview (Speaking) Scores



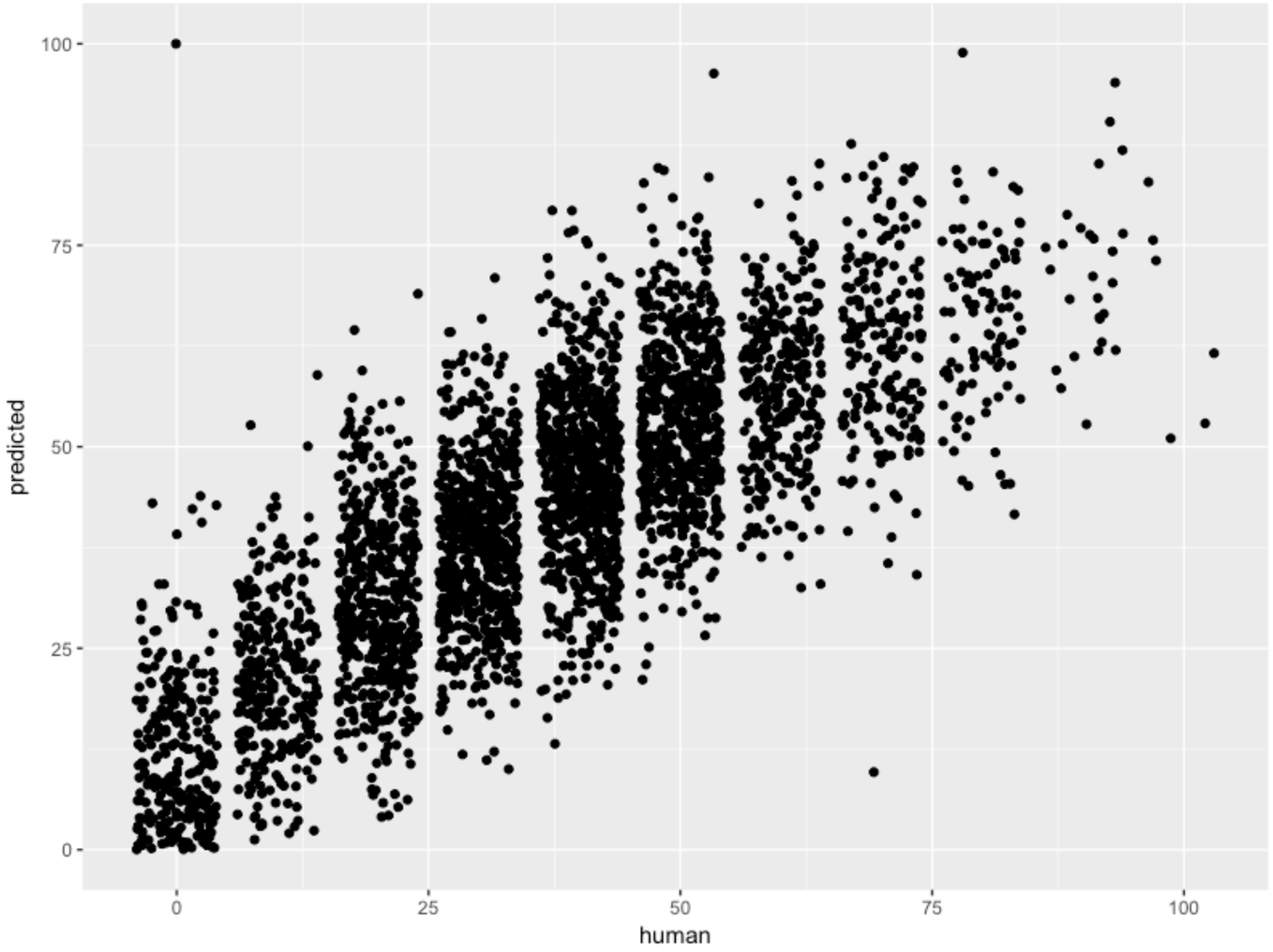
human-machine agreement: **0.79**

human-human agreement: **0.77**

reliability: **0.91**

duolingo english test

Automated Writing Scores



human-machine agreement: **0.82**

human-human agreement: **0.68**

reliability: **0.80**

duolingo english test

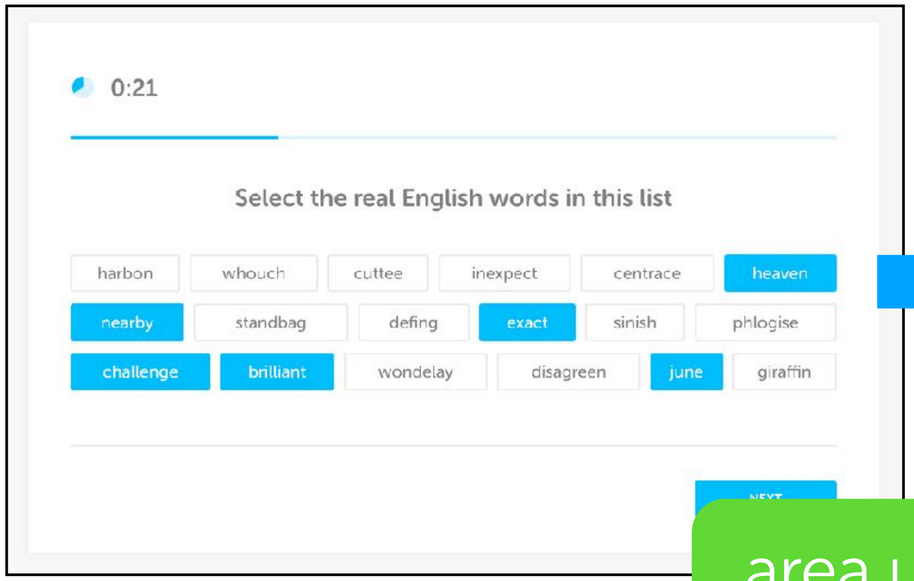
Projecting the Score Level

- use a **regression** to predict CEFR-derived score level as a function of:
  - statistics from an English **Markov chain model** (a.k.a. the “Fisher score”)
  - **word length** (characters, syllables)
- note: these features are available for **pseudowords**, as well!

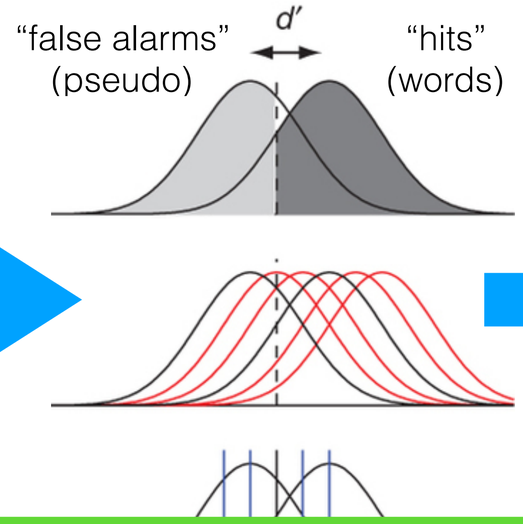
Grading & Generalized IRT

(McNichol, 1972; Milton, 2010; Wickens, 2002; Zimmerman et al., 1977)

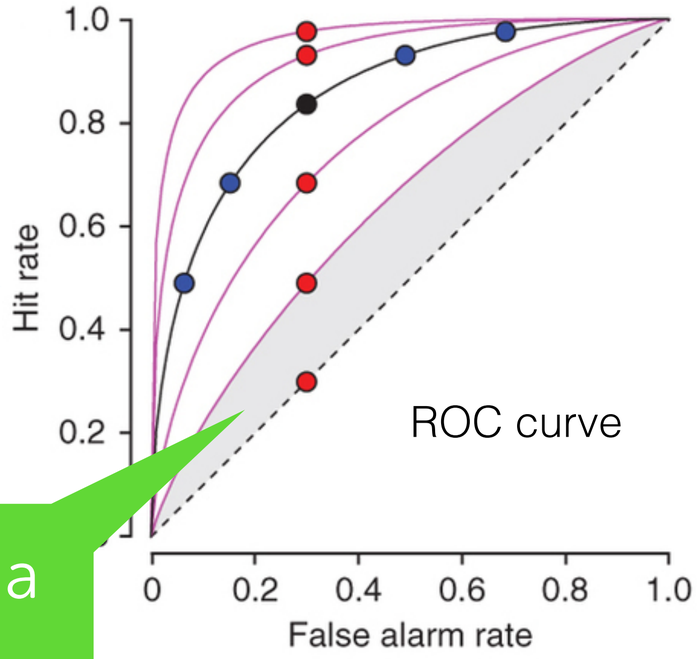
determine "hits," "false alarms," etc.



SDT analysis



ROC analysis

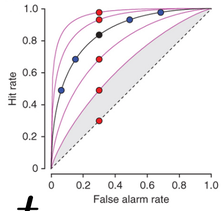


area under the ROC curve has a **probabilistic** interpretation

Grading & Generalized IRT

(c.f., Sands et al., 1997; Segall, 2005)

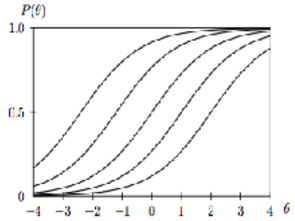
probabilistic item/testlet grades can be **directly employed** into an IRT-based CAT framework



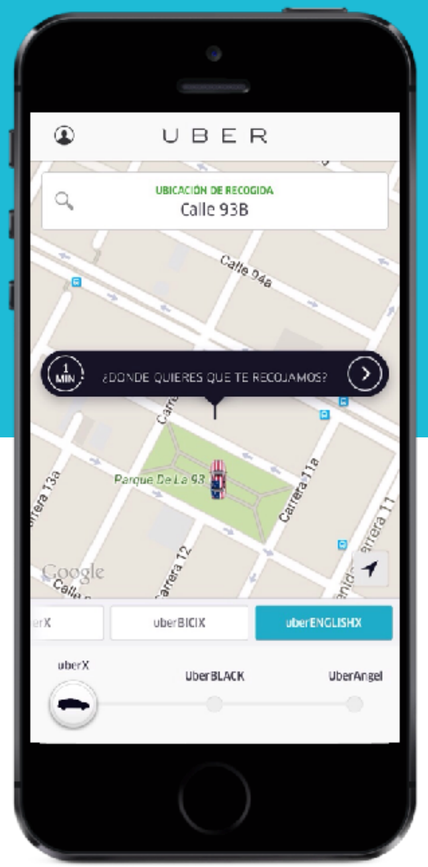
probabilistic grade
(SDT/ROC; in range [0-1])

$$\hat{\theta}_t^* = \arg\max_{\theta} \log \prod_{i=1}^{t} p_i(\theta)^{g_i} (1 - p_i(\theta))^{1-g_i}$$

item response function prediction
(Rasch model; $\mathbf{b_i}$ is CEFR-derived score level)



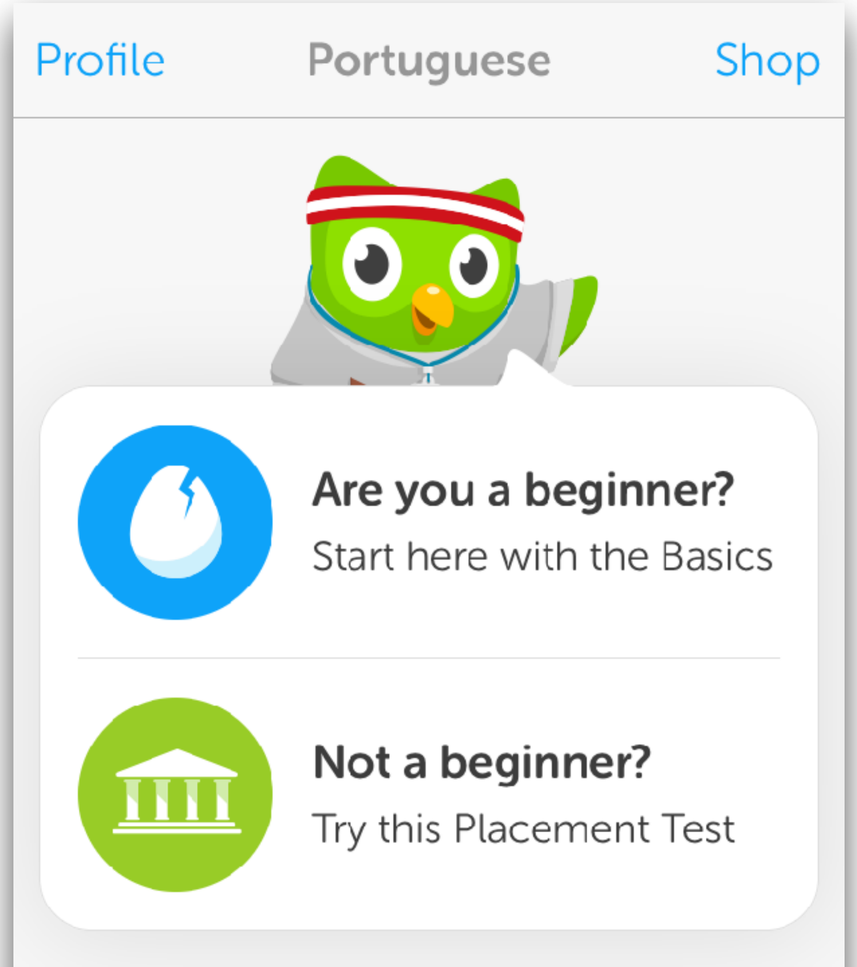
Other Use Cases



UberENGLISH uses DET to certify English skills for their drivers

Business Adoption

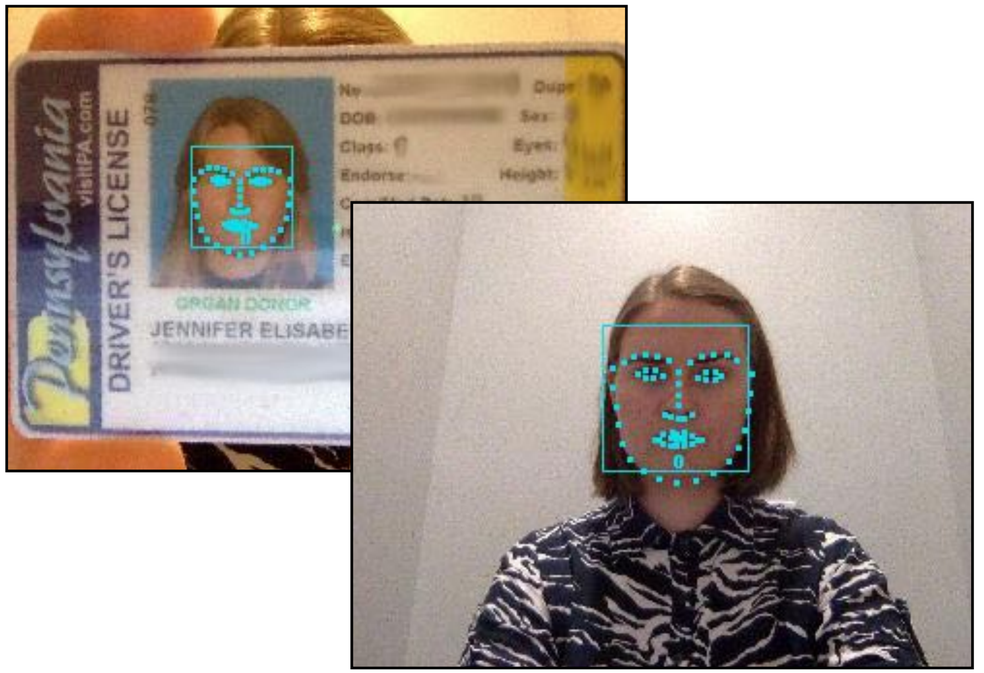
The Beginning:
Duolingo's Placement Test



Impact & Consequences

Security

- we're using **data forensics** to help verify examinee identity & test integrity
  - **face** detection & recognition
  - **voice** recognition
  - **keystroke/mouse** biometrics
  - **eye** tracking, **lip** syncing



"Thank you for helping us with students applying to us through a new initiative called the **Scholarship Program for Displaced Persons**, designed to aid prospective students impacted by the Syrian civil war."

COLUMBIA | GS
School of General Studies

"[We had] a case last year involving a recruited **athlete** in Eastern Europe that the coach found late in the day and who **couldn't schedule other testing**. And I know the DET was important to the admission decision."

Yale University

"I think there is great potential for geographic **diversity within countries** as a result of the DET. One example would be **Brazil**, which is the fifth largest country in the world, and has a rapidly growing number of students applying to the US.

"Recently there was a **trucker's strike** that shut down transportation throughout the country for over a week… a lot of talented students had to **miss their testing opportunity**. My guess is there are states of Brazil where the nearest TOEFL center is many **hours away** assuming transportation is even running. The DET really opens up access because of its ability to be taken on demand."

Dartmouth