East Asia Assessment
Solutions Team

AI and Language Assessment: Challenges and Potential

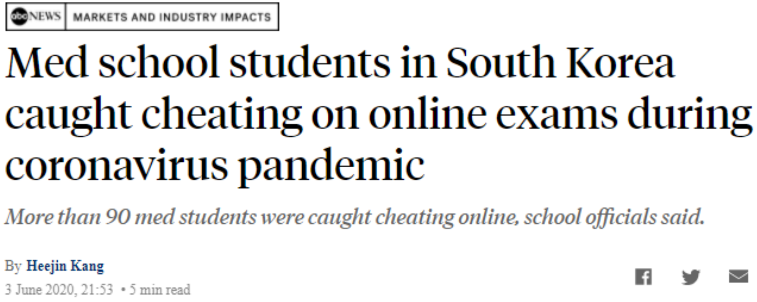
International Virtual Conference on TESOL
Vietnam, 27 November 2020

Sheryl Cooke
Trevor Breakspear

BRITISH COUNCIL

East Asia Assessment
Solutions Team

Computers and language testing:
Do you trust online testing?
Do you trust machines to give an accurate score?

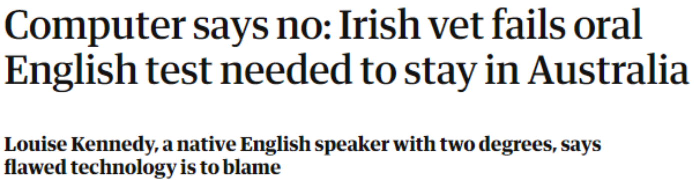


abc NEWS | MARKETS AND INDUSTRY IMPACTS

Med school students in South Korea caught cheating on online exams during coronavirus pandemic

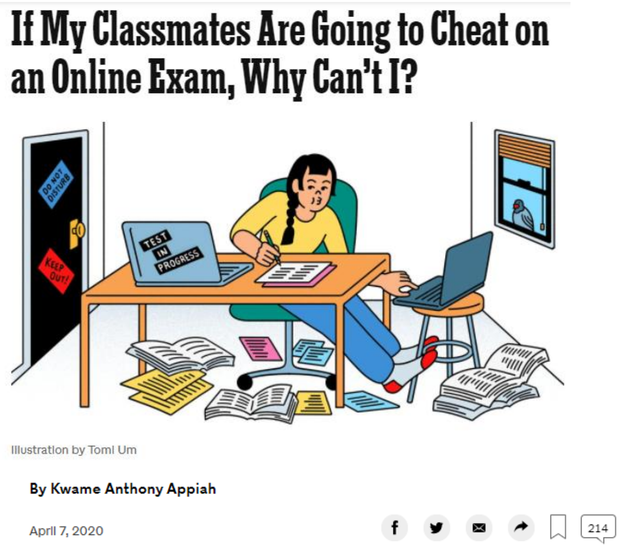
More than 90 med students were caught cheating online, school officials said.

By Heejin Kang
3 June 2020, 21:53 • 5 min read



Computer says no: Irish vet fails oral English test needed to stay in Australia

Louise Kennedy, a native English speaker with two degrees, says flawed technology is to blame



If My Classmates Are Going to Cheat on an Online Exam, Why Can't I?

Illustration by Tomi Um

By Kwame Anthony Appiah

April 7, 2020

214

What is possible?

What are the current (or near-future) capabilities of technology in language assessment?

What is feasible?

What is realistically achievable for most organisations and educational institutions?

What is desirable?

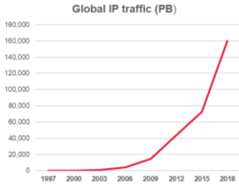
What would be good – or bad – for test-takers and other stakeholders?

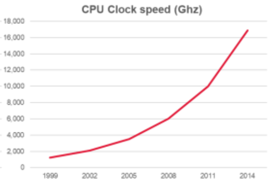
Drivers

- Scale
- Economy
- Access
- Speed
- Social distancing

Enablers

- Computer processing power
- Internet development
- Big data and machine learning
- Computer literacy
- Automated Speech Recognition (ASR)





The influence of technology on language assessment

Delivery

- All modules, or only some
- With a human interlocutor, or only the computer
- In a centre, or remotely

Feedback

- Writing
- Speaking

Rating

- Machine only
- Hybrids:
  - Different criteria
  - Mostly machine, some human
  - Mostly human, some machine

Delivery

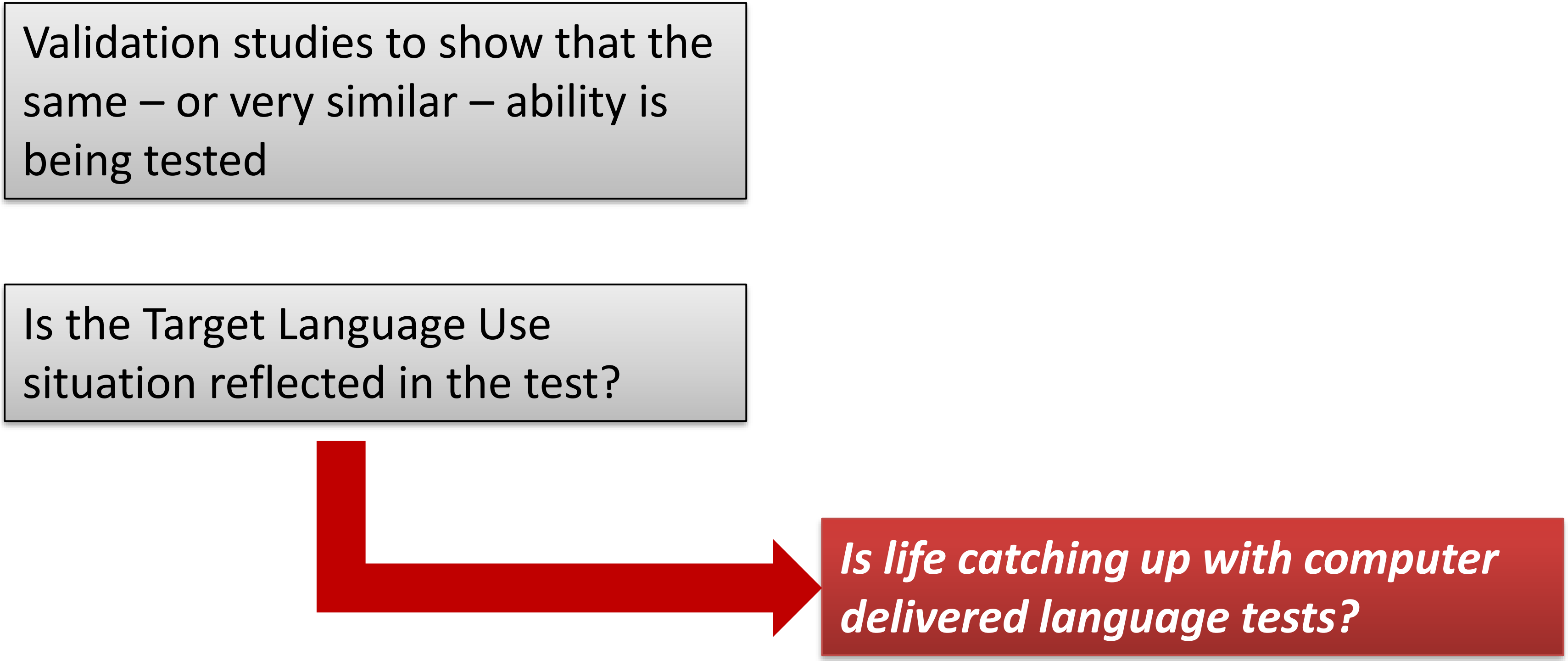
Computers for test delivery

✓ Possible

✓ Feasible

? Desirable →

- Are there any unintended consequences or impact?
- Does the medium of delivery change the construct?
  - Is a computer-delivered test still testing the same abilities?
  - Is the test still measuring the same thing?
- Security



Validation studies to show that the same – or very similar – ability is being tested

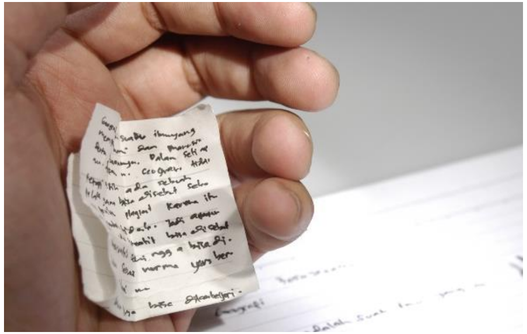
Is the Target Language Use situation reflected in the test?

Is life catching up with computer delivered language tests?

Test security: Remote proctoring

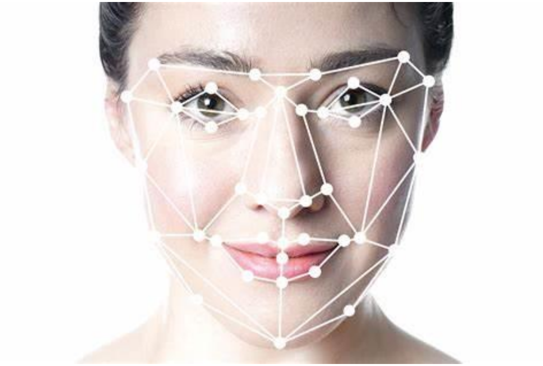
Problems:

- Over 70% college students reported cheating during postsecondary education (Whitley, 1998)
- Students cheat to get ahead (Simkin and McLeod, 2010)
- Cheating is easier online because:
  1. It can be more difficult to authenticate identity
  2. It is more difficult to monitor behavior (camera coverage/ proctor resourcing)

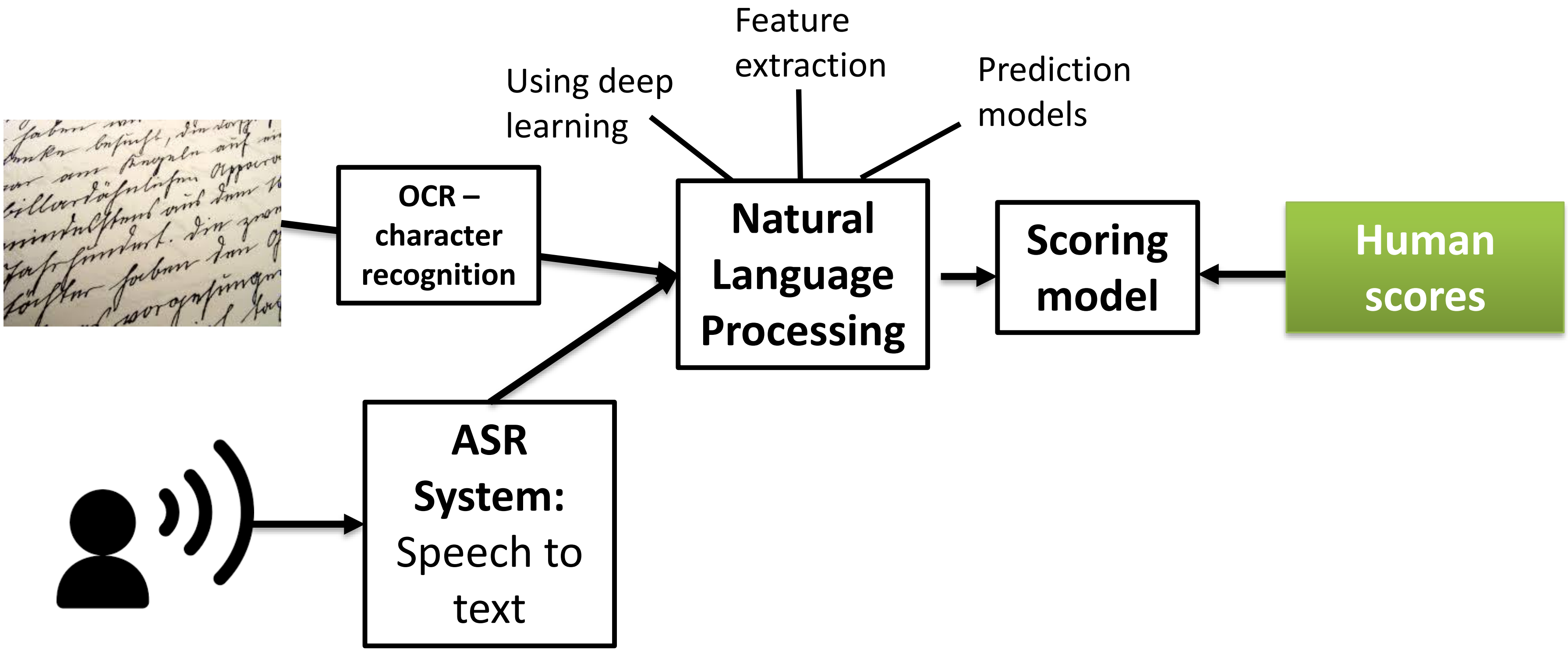


Solutions?

- AI-trained image recognition systems used to validate authenticity
- Voice and sound recognition system used to identify suspicious background noise (multiple voices etc.)
- AI-trained facial recognition systems used to identify signs of cheating
- Suspicious examples sent to a human proctor for further investigation



Rating



Feature extraction
Using deep learning
Prediction models
OCR – character recognition
Natural Language Processing
Scoring model
Human scores
ASR System: Speech to text

Computers for rating productive skills

✓ Possible →
- With limitations!!!

✓ Feasible →
- Large amounts of data
- Continuous validation

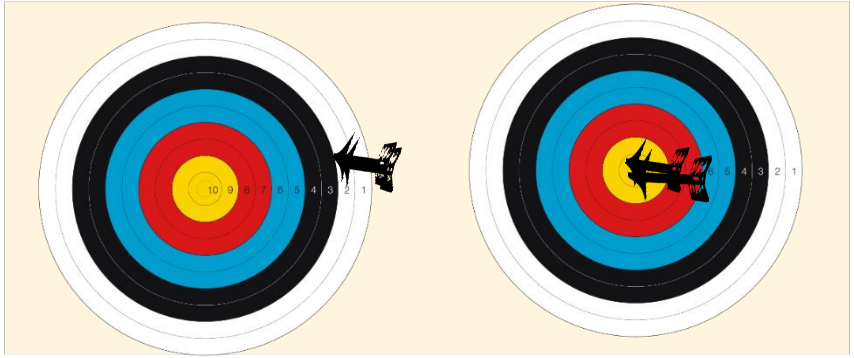
? Desirable

Automated assessment of productive skills: Key concerns and considerations

- Correlation ≠ validation
- Narrowing of the construct
- Bias

A reliable but invalid test.

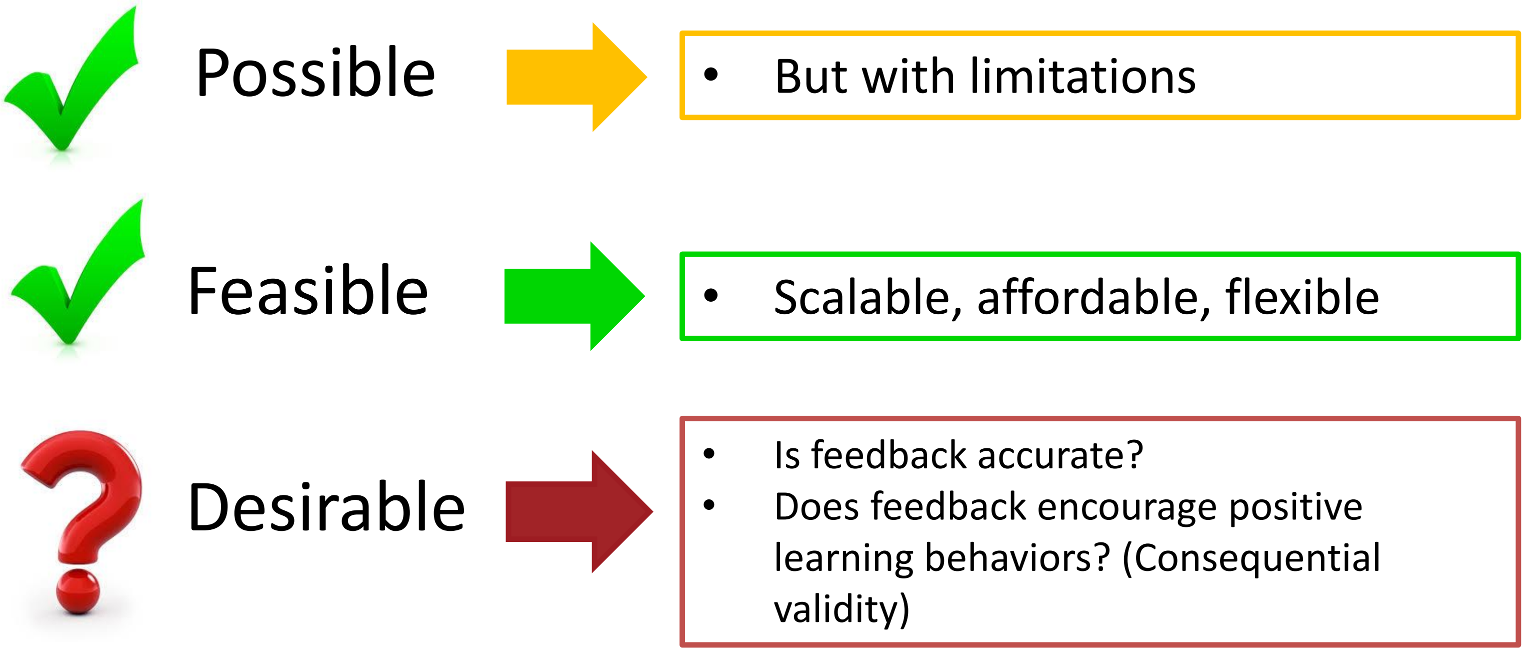
A valid test



> Just a new tool, one that can be used for good and for bad purposes. We know already that machine learning has huge potential, but data sets with biases will produce biased results - garbage in, garbage out.
>
> *Sarah Jeong, Journalist specializing in IT law*

Feedback

Feedback

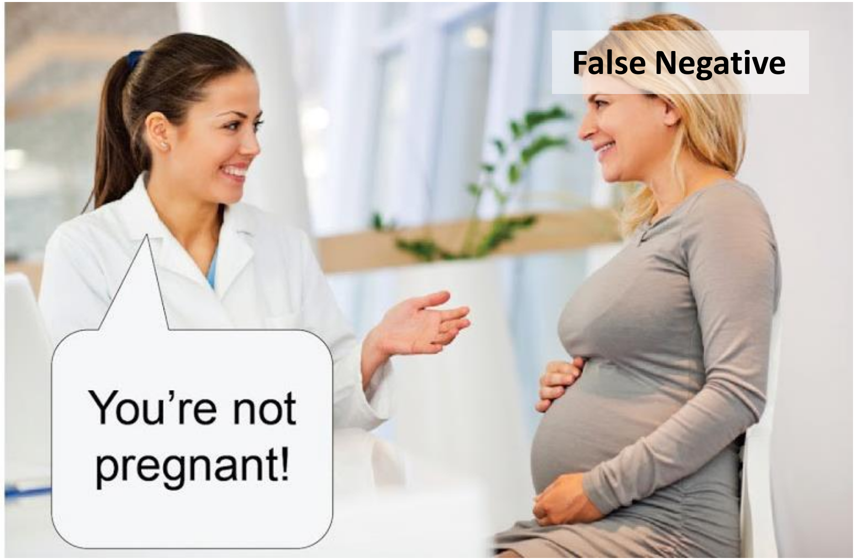


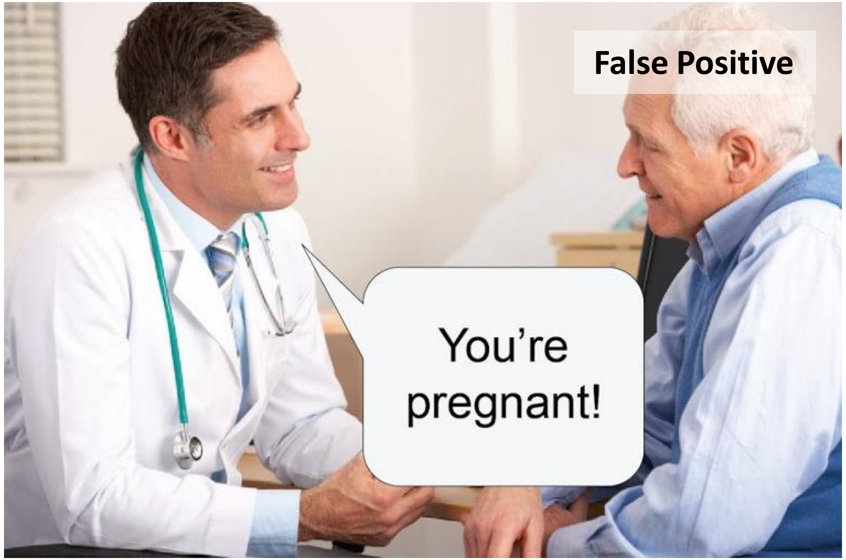
✓ Possible →
- But with limitations

✓ Feasible →
- Scalable, affordable, flexible

? Desirable →
- Is feedback accurate?
- Does feedback encourage positive learning behaviors? (Consequential validity)

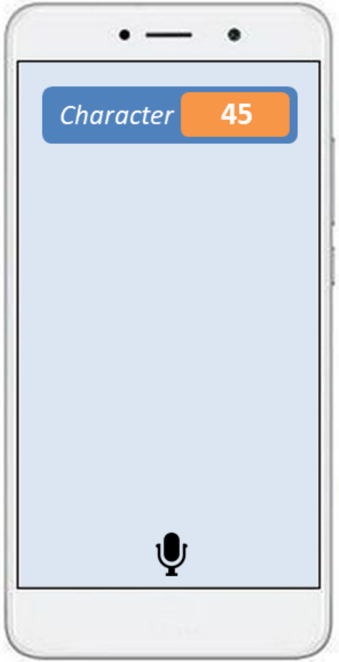
AI concepts: false positive or false negative?





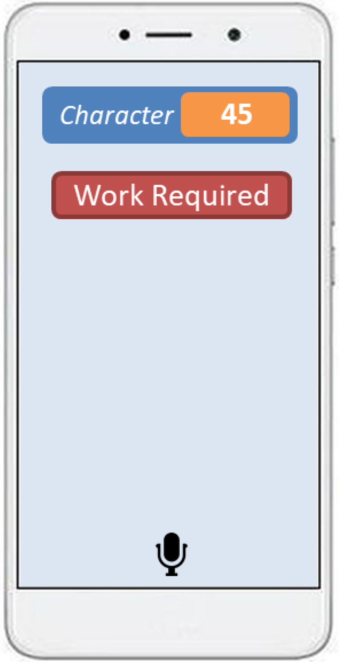
Critical analysis: Pronunciation feedback

Word-level Feedback Parameters			
Overall Score			
1-100			



Critical analysis: Pronunciation feedback

Word-level Feedback Parameters			
Overall Score	**Type of performance**		
1-100	1. Highlights 2. Work required		

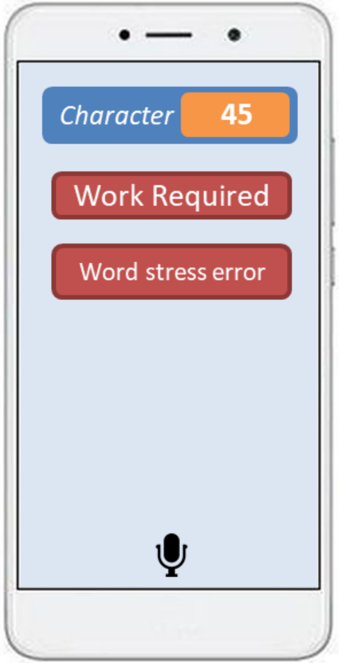


Character 45

Work Required

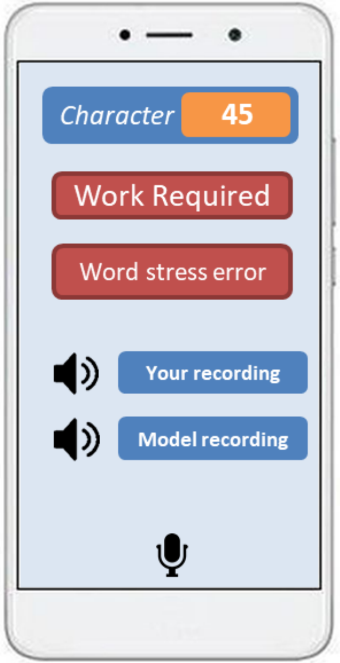
Critical analysis: Pronunciation feedback

Word-level Feedback Parameters			
Overall Score	**Type of performance**	**Error category**	
1-100	1. Highlights 2. Work required	1. Phonetic accuracy 2. Word stress	

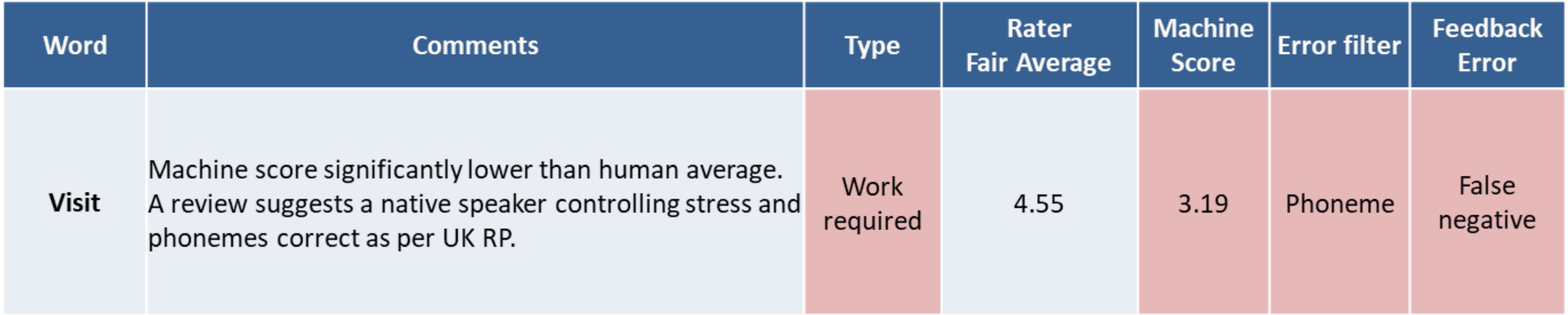


Critical analysis: Pronunciation feedback

Word-level Feedback Parameters			
Overall Score	**Type of performance**	**Error category**	**Recordings for learner**
1-100	1. Highlights 2. Work required	1. Phonetic accuracy 2. Word stress	1. Model answer 2. Learner answer



Fit for purpose: Pronunciation feedback



Word	Comments	Type	Rater Fair Average	Machine Score	Error filter	Feedback Error
Visit	Machine score significantly lower than human average. A review suggests a native speaker controlling stress and phonemes correct as per UK RP.	Work required	4.55	3.19	Phoneme	False negative

Fit for purpose: Pronunciation feedback

Word	Comments	Type	Rater Fair Average	Machine Score	Error filter	Feedback Error
Visit	Machine score significantly lower than human average. A review suggests a native speaker controlling stress and phonemes correct as per UK RP.	Work required	4.55	3.19	Phoneme	False negative
character	Machine score significantly higher than human average. A review shows that the second syllable is incorrectly stressed which could explain the overrating. This finding is supported by the Machine syllable stress error filter result.	Work required	3.40	4.96	Stress	False positive

So, where does this leave us with technology?

Language testing is at a moment of crisis [as a result of technology and sociolinguistic factors converging].

Tim McNamara, Language Assessment Researcher

Careful evaluation of the consequences

Leveraging the best of both humans & technology – hybrid approaches

The End

References

https://abcnews.go.com/International/med-school-students-south-korea-caught-cheating-online/story?id=71043491

https://www.nytimes.com/2020/04/07/magazine/if-my-classmates-are-going-to-cheat-on-an-online-exam-why-cant-i.html

https://www.theguardian.com/australia-news/2017/aug/08/computer-says-no-irish-vet-fails-oral-english-test-needed-to-stay-in-australia

Thank You!

Any Questions?

Sheryl.cooke@britishcouncil.org.cn

trevorjohn.breakspear@britishcouncil.org.cn

https://www.britishcouncil.cn/en/exams/EAAST