BRITISH COUNCIL

East Asia Assessment Solutions Team

Validating the use of autorating technologies in the assessment of speaking skills

Trevor Breakspear

Jan Langeslag, William Bayliss, Johnathan Cruise, Frank Wucinski

BRITISH COUNCIL

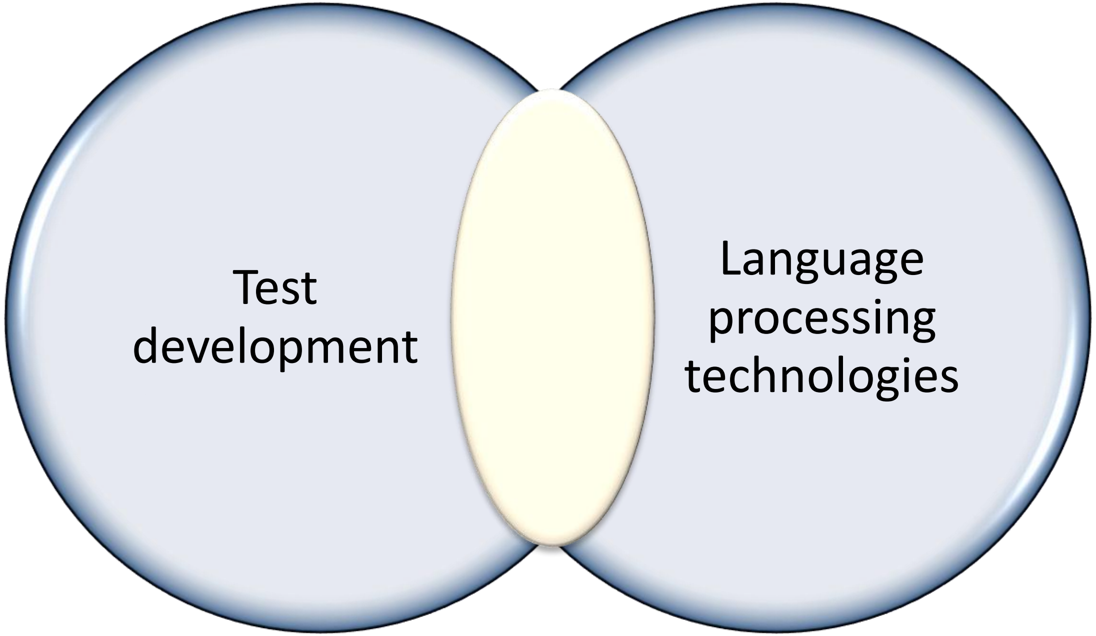
What is AI?

AI is not a he or a she or even an it, AI is more like a “they.”
Rob Smith, CEO of Pecabu

By far, the greatest danger of Artificial Intelligence is to conclude too early that we understand it.
Eliezer Yudkowsky, AI Researcher

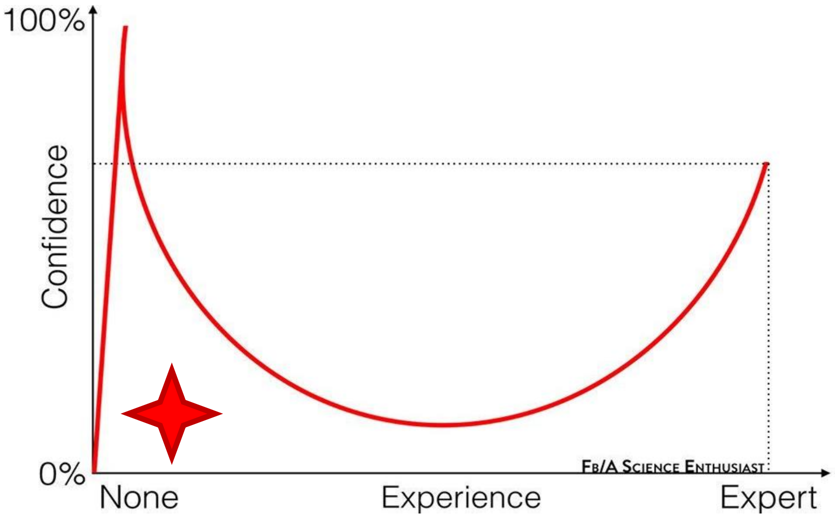
We know already that machine learning has huge potential, but data sets with biases will produce biased results - garbage in, garbage out.
Sarah Jeong, Journalist specializing in IT law

Where we want to be

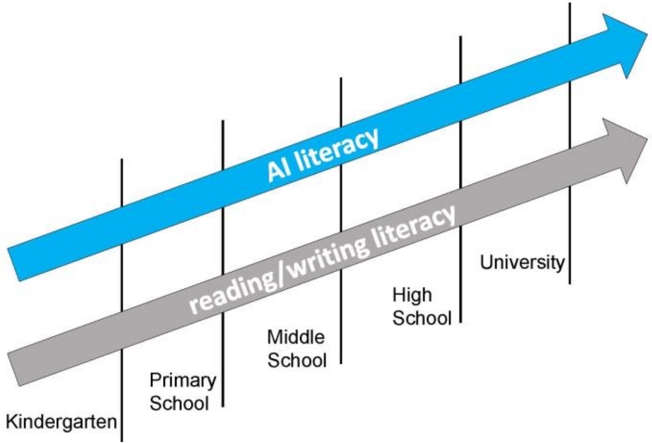


"..the transdisciplinary space at the intersection of language processing technologies and second language assessment." (Chapelle & Chung, 2010)

Where I am in Data Science!



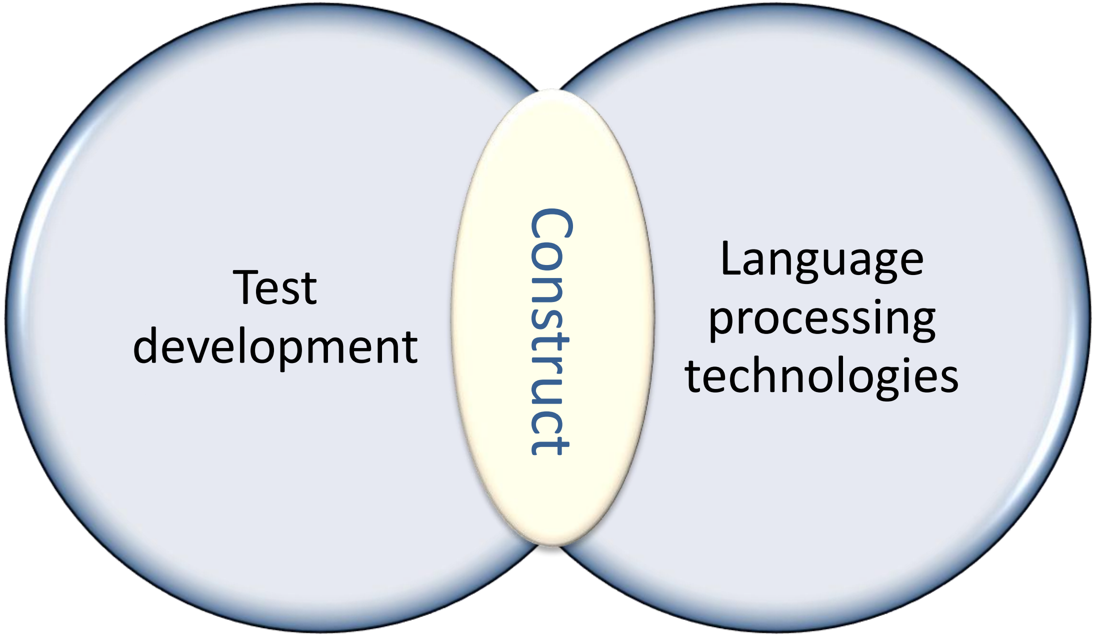
https://ritholtz.com/wp-content/uploads/2018/04/DbkaOnJV0AA9dWb.jpg



Kandlhofer et al. 2016

BRITISH COUNCIL

Alignment



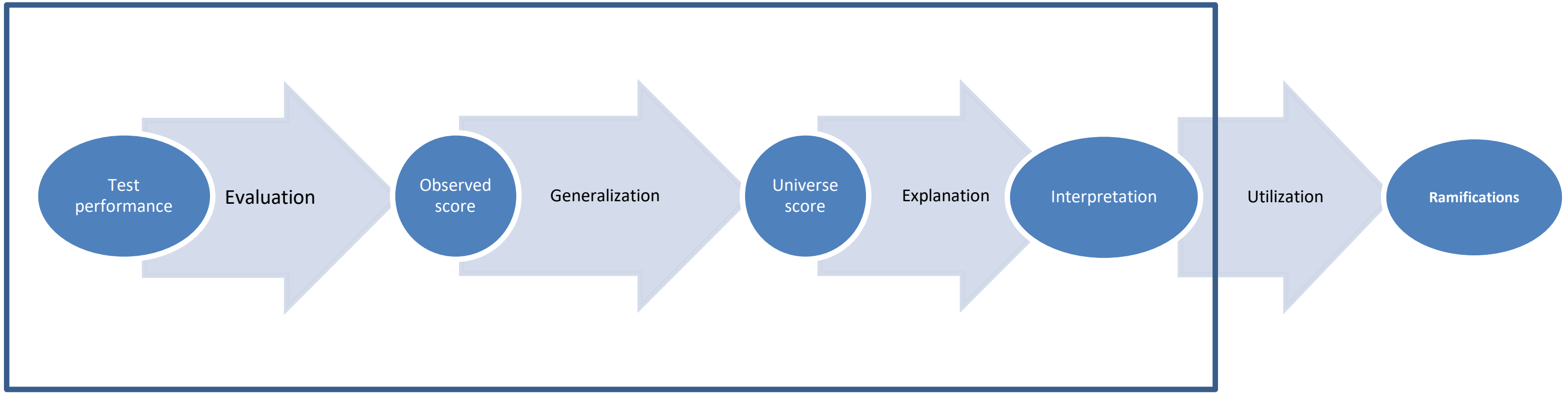
Fit for purpose: align the technical potential and best practice test development; minimize the limitations

BRITISH COUNCIL

Validation approach

Inferential validity argument based on supporting assumptions (decisions made) at key stages of the test development process (Kane, 1992)

- What are the difference between human and autorated assumptions and implications?

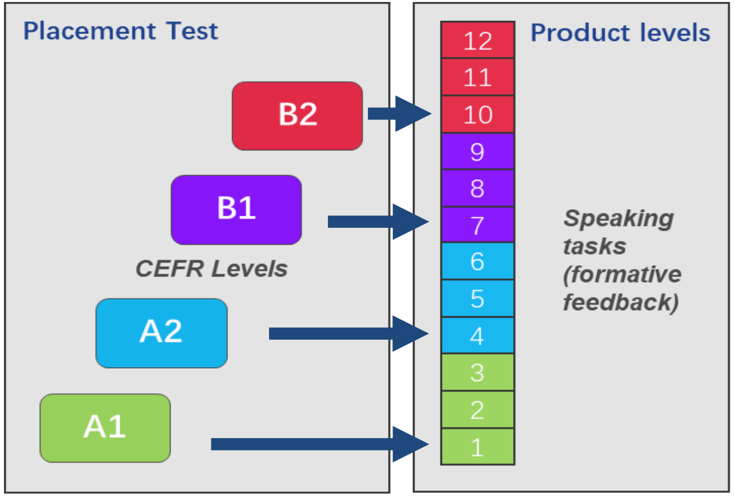


BRITISH COUNCIL

Assessment tool overview

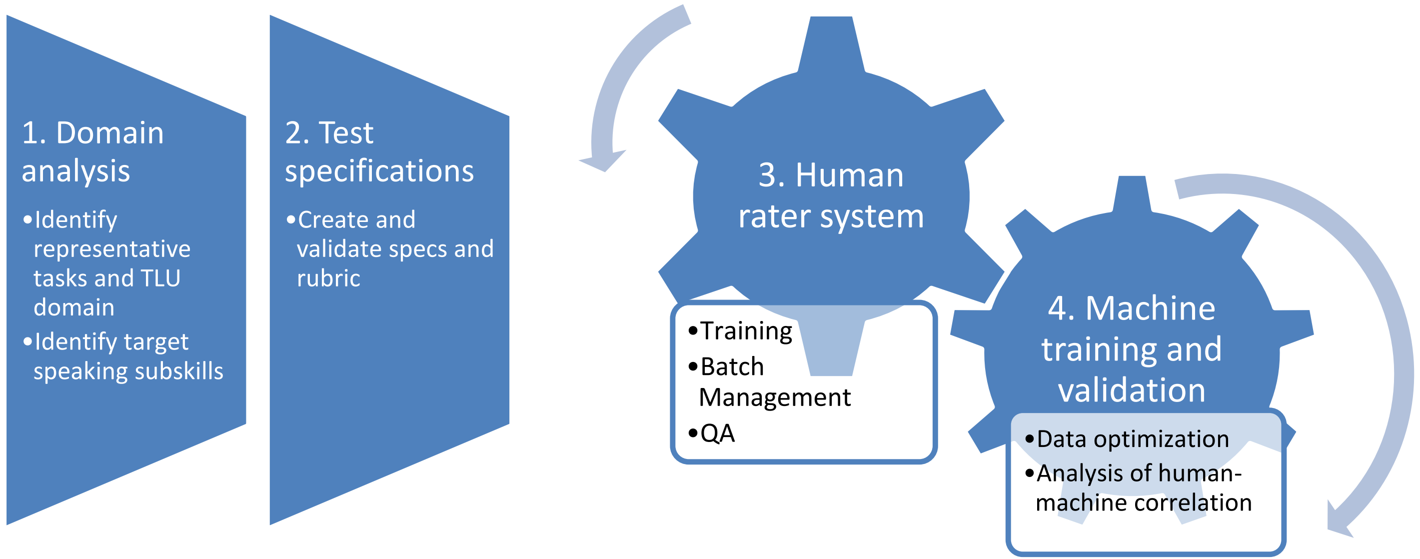
IELTS Smart Learning (ISL) owned by the IELTS Partners

- Low-stakes AI rated speaking solution for Chinese students aged 12-16, CEFR level A1-B2.
- Placement test includes constrained and open-ended response tasks
- Placement test is holistically rated and provides recommended product level to start practice.
- Scoring rubrics and item specifications based on CEFR
- Formative content includes activities over 12 product levels with AI enabled scores and feedback.



BRITISH COUNCIL

Autorater training overview



- ✓ Consistent group of 6 gold-standard raters trained engine over 6-month period on a 0-6 scale (A1-B2+)
- ✓ Each candidate response scored by between 3 and 6 raters
- ✓ Statistical modelling techniques, regular training and weekly rater feedback improve training data reliability

Scoring assumptions (1)

Human rating

Evaluation

Observations are targeting features defined in the construct

Machine rating

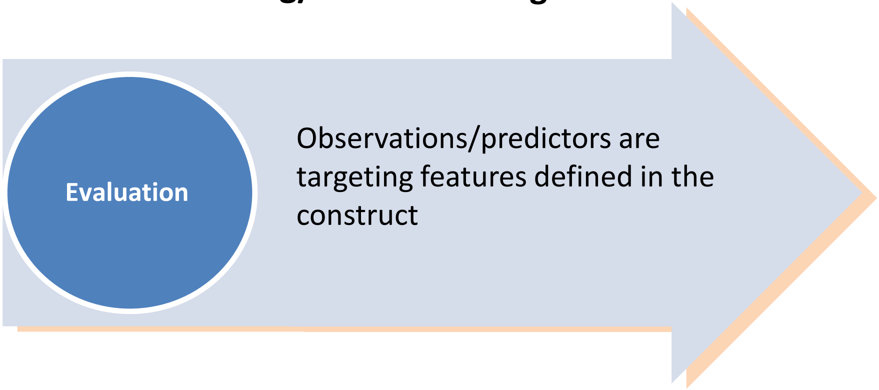
Evaluation

Machine proxies are targeting features defined in the construct

BRITISH COUNCIL

Scoring assumptions (1)

Human rating/Machine rating correlation



Commonly expressed as a correlation between machine and human ratings (where 0 = no correlation, and 1 = 100% correlation)

Scoring assumptions (1)

Machine rating **Human rating**

BUT

1. Are human ratings targeting the speaking features required?
2. Do machine predictors equate to features used by humans?
3. Consequences of construct underrepresentation (gaming/unconventional responses)

Scoring assumptions (2)

Human rating

Generalization

Scores are consistent across different test forms, test taker groups

Machine rating

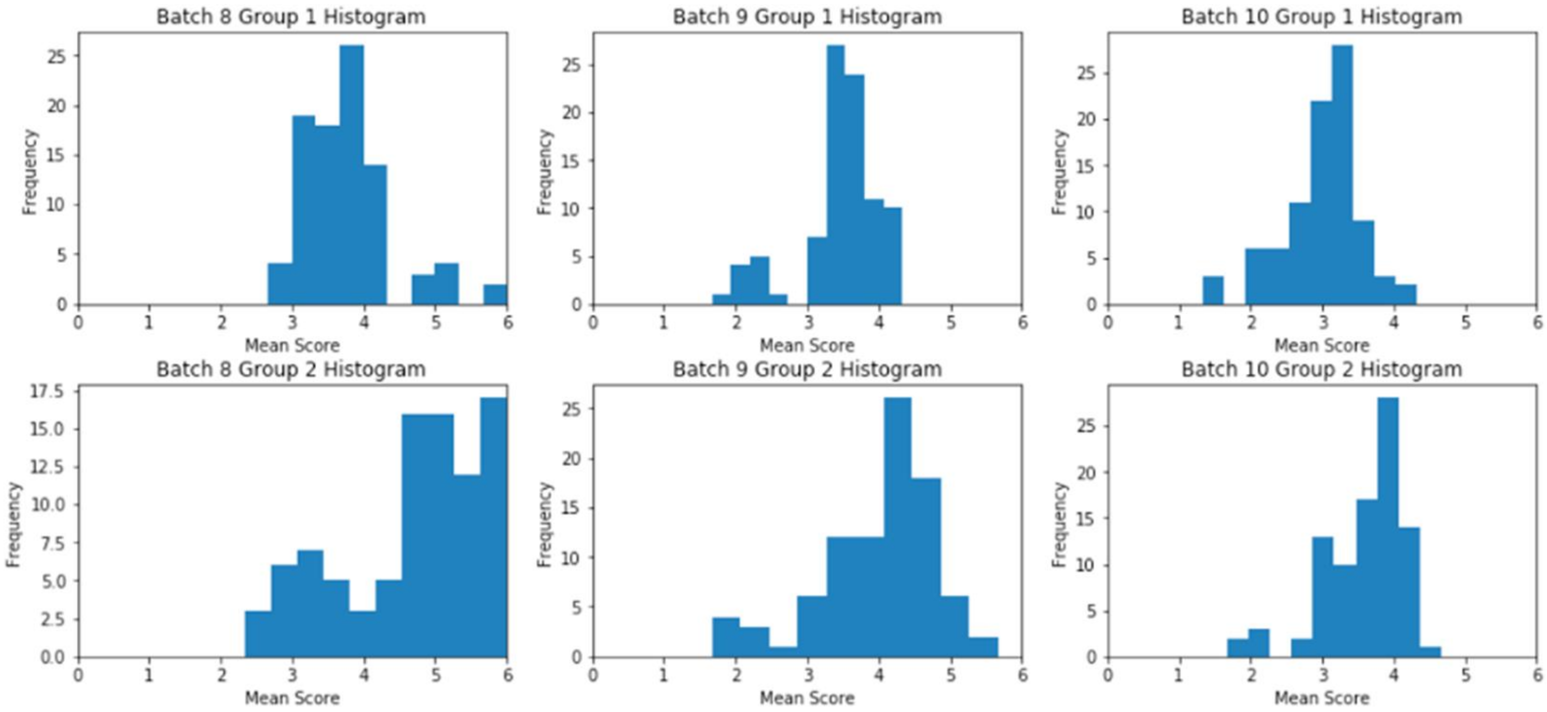
Generalization

Representative sample used to train the model

How generalizable is the scoring model to the general test taker population?

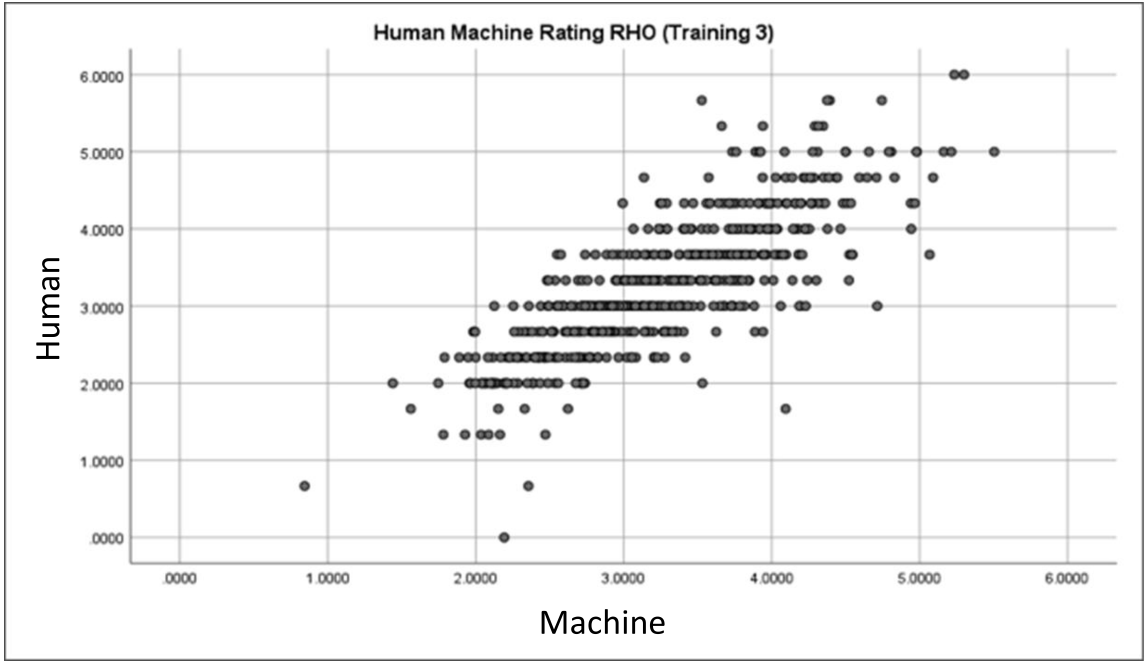
BRITISH COUNCIL

Original proficiency distribution



BRITISH COUNCIL

Human-machine correlation (3)

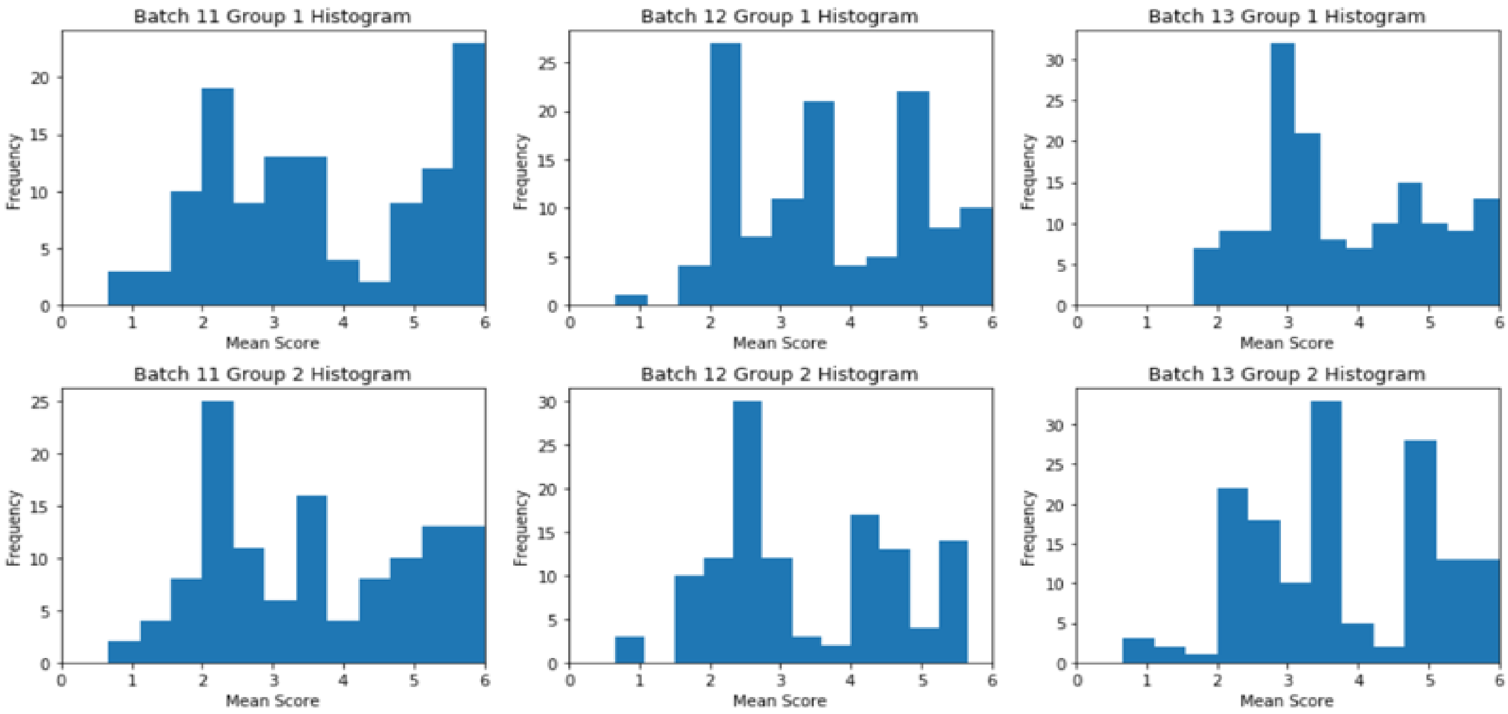


Training 3	
RHO Correlation Coefficient	.803**

**. Correlation is significant at the 0.01 level (2-tailed).

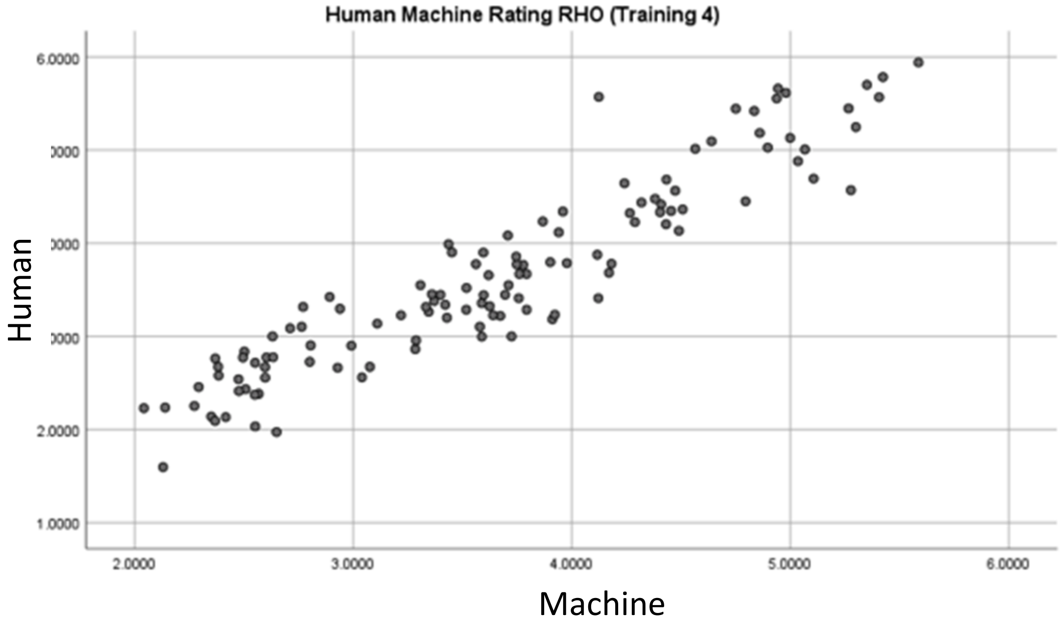
BRITISH COUNCIL

Enhancing sample distribution



BRITISH COUNCIL

Human-machine correlation (4)



Training 4	
RHO Correlation Coefficient	.93**
**. Correlation is significant at the 0.01 level (2-tailed).	

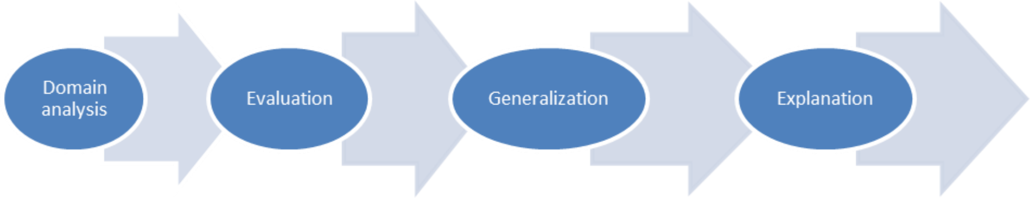
In-development validation

1. Are human ratings targeting the speaking features required?

2. Do machine predictors equate to features used by humans?

3. Consequences of construct underrepresentation (gaming/unconventional responses)

- **Iterative development = iterative validation**



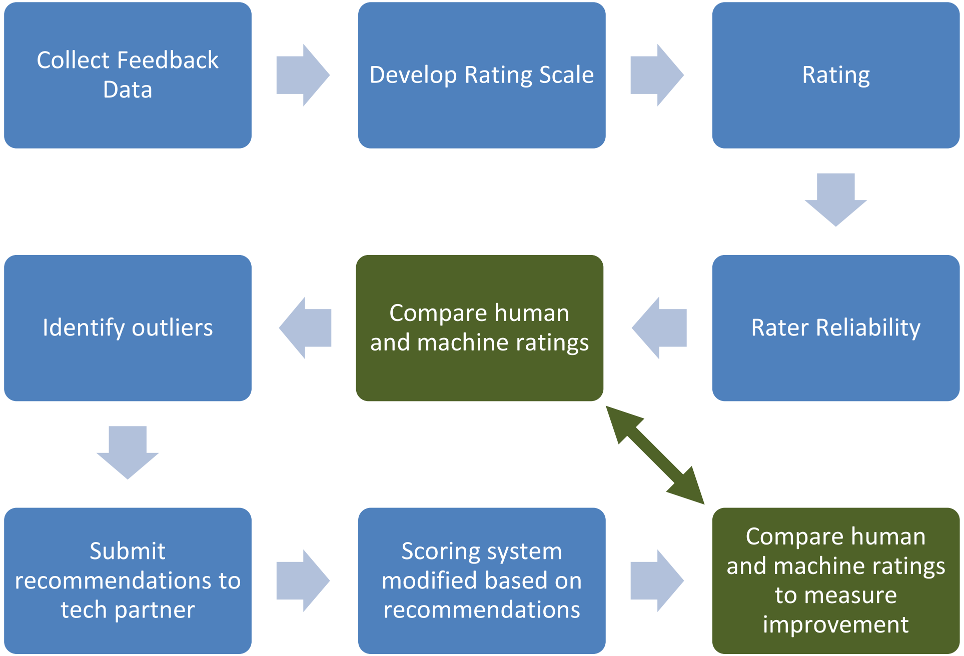
In-development validation

1. Are human ratings targeting the speaking features required?
Discourse analysis
2. Do machine predictors equate to features used by humans?
Larger representative data sample to train – work needed
3. Consequences of construct underrepresentation (gaming/unconventional responses)
Outlier analysis/extraction filters – more work needed

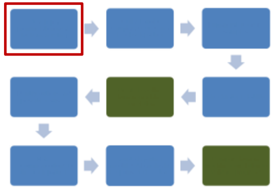
- **Iterative development = iterative validation**

Word level pronunciation validation: Purpose and Approach

- To consider the efficacy of pronunciation scores delivered by an AI engine at the word level
- To provide meaningful feedback to tech partner on weaknesses of the current AI feedback system
- To facilitate improvement in the quality of AI-driven pronunciation feedback to learners



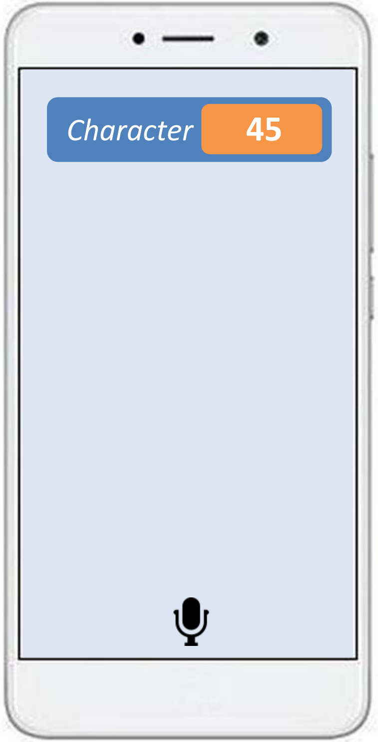
BRITISH COUNCIL

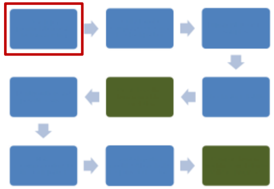


East Asia Assessment
Solutions Team

Collect Feedback Data

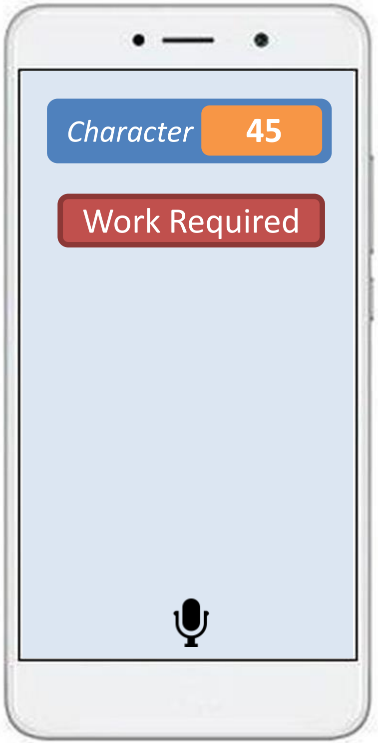
Word-level Feedback Parameters			
Overall Score			
1-100			

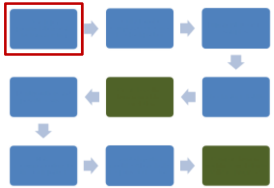




Collect Feedback Data

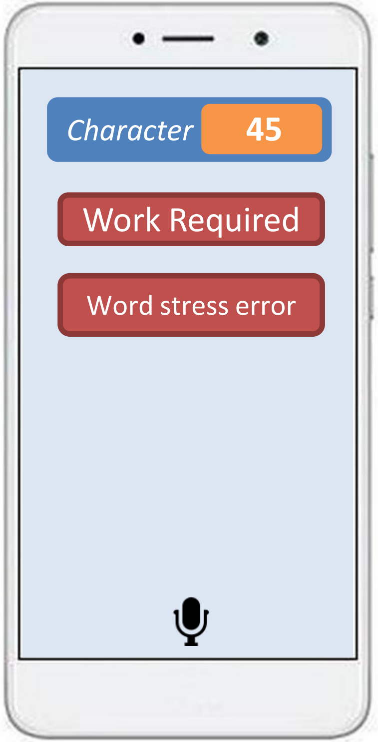
Word-level Feedback Parameters			
Overall Score	**Type of performance**		
1-100	1. Highlights 2. Work required		



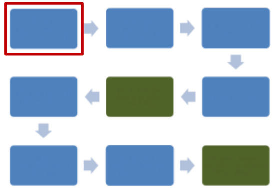


Collect Feedback Data

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

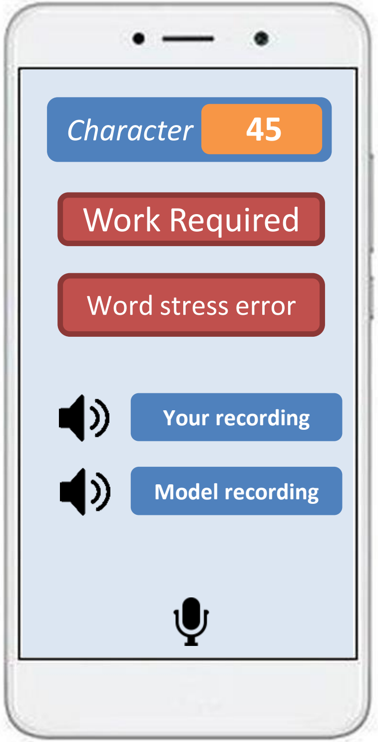


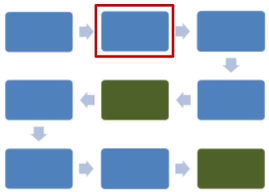
East Asia Assessment
Solutions Team



Collect Feedback Data

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

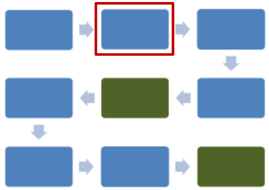




Developing the word-level rating scale

- Development based on three rating scale principles (1) clear (2) concise and (3) discrete
- Scale based on the pronunciation features (1) **phoneme accuracy**

Descriptor	Score
The word was produced as commonly pronounced in recognized global English variants.	5
The word was produced as commonly pronounced in any recognized global English variant, minor issues with syllable delivery	4
Lapses in phonetic accuracy may be noticeable	3
Phoneme delivery may be faulty	2
Mispronounced phoneme, or intrusive substitution or deletion of a phoneme	1

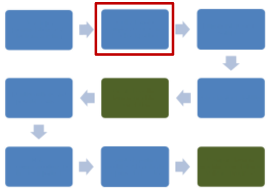


Developing the word-level rating scale

- Development based on three rating scale principles (1) clear (2) concise and (3) discrete
- Scale based on the pronunciation features (1) **phoneme accuracy** (2) **word stress**

Descriptor	Score
The word was produced as commonly pronounced in recognized global English variants. Appropriate number of syllables and **stress was placed on the correct syllable**.	5
The word was produced as commonly pronounced in any recognized global English variant. Appropriate number of syllables and stress placed on the correct syllable. There may be **minor issues with syllable** or **stress** delivery.	4
Lapses in phonetic accuracy and/or word stress	3
Phoneme delivery and/or **stress may both be faulty**	2
mispronounced phoneme, or intrusive substitution or deletion of a phoneme and stress may be faulty.	1

BRITISH COUNCIL



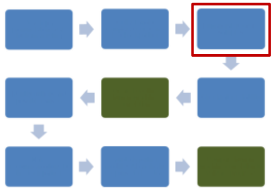
East Asia Assessment
Solutions Team

Developing the word-level rating scale

- Development based on three rating scale principles (1) clear (2) concise and (3) discrete
- Scale based on the pronunciation features (1) **phoneme accuracy** (2) **word stress accuracy** and (3) **effect on intelligibility**
- Results of rater perception survey showed examiners considered descriptors clear and easy to use (4.33/5)

Descriptor	Score
The word was produced as commonly pronounced in recognized global English variants. Appropriate number of syllables and **stress was placed on the correct syllable**.	5
The word was produced as commonly pronounced in any recognized global English variant. Appropriate number of syllables and stress placed on the correct syllable. There may be **minor issues with syllable** or **stress** delivery.	4
The word as produced is intelligible. **Lapses in phonetic accuracy and/or word stress may be noticeable but cause little strain.**	3
The word as produced is barely intelligible. **Phoneme delivery** and/or **stress may both be faulty causing some strain and possible misunderstanding**.	2
The word produced may include a **mispronounced phoneme, or intrusive substitution or deletion of a phoneme and stress may be faulty**. **Resulting in misunderstanding or the pronunciation of a different word.**	1

East Asia Assessment
Solutions Team



Rating

- Five experienced examiners chosen to provide ratings, training provided and feedback given on unexpected responses (from FACETS) - Any outlier samples sent back for rater re-analysis
- The fair average score from FACETS software used as final score
- Total of 280 word samples rated, 109 **work required**, 171 **highlights**

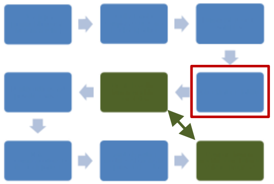
Process

Listen to mp3 sample → Transcribe → Listen to model answer → Provide a score using the rating scale

Rating Sheet

Item No.	Learning recording link	Transcribe	Model answer link	Score
1	learner1.com	xxxx	model1.com	3
2	learner2.com	xxxx	model2.com	4
3	learner3.com	xxxx	model3.com	5

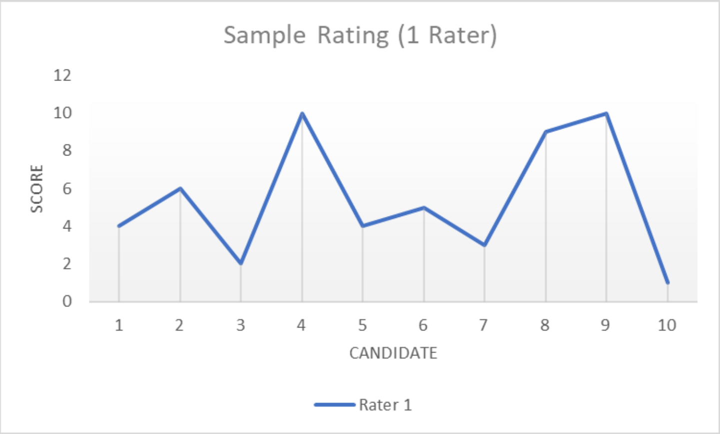
BRITISH COUNCIL



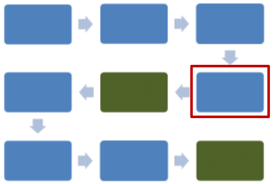
East Asia Assessment
Solutions Team

Rater Reliability

- FACETS software used to calculate the degree of agreement between rater scores (interrater reliability)
- Three main aspects of reliability were considered, **separation** (leniency severity) **fit,** and **exact agreements**
- These three principles can be illustrated using the following (fictional) example:



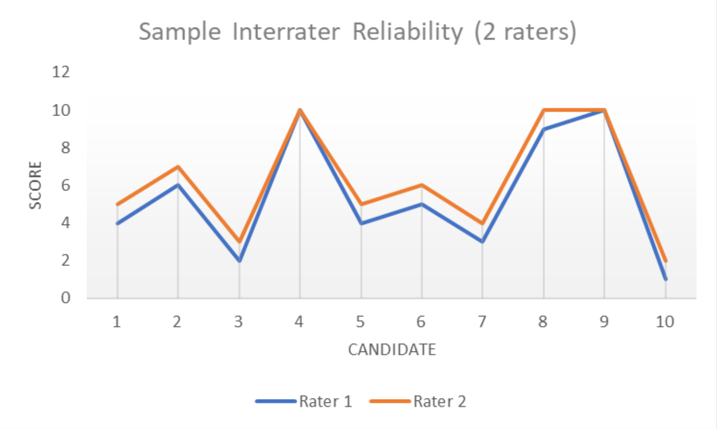
East Asia Assessment
Solutions Team

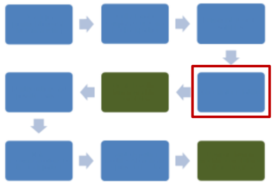


Rater Reliability

- FACETS software used to calculate the degree of agreement between rater scores (interrater reliability)
- Three main reliability factors were considered, **separation** (leniency severity) **fit,** and **exact agreements**
- These three principles can be illustrated using the following (fictional) example:

- *Rater 1 is generally more severe than rater 2*
- *Rater 1 and 2 show exact agreement agree for candidates 4+9*

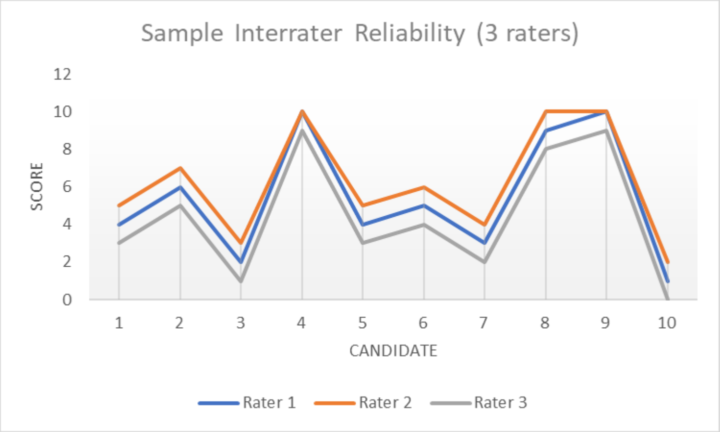




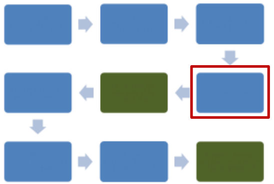
Rater Reliability

- FACETS software used to calculate the degree of agreement between rater scores (interrater reliability)
- Three main reliability factors were considered, **separation** (leniency severity) **fit,** and **exact agreements** These three principles can be illustrated using the following (fictional) example:

- *Rater 1 is generally more severe than rater 2*
- *Rater 1 and 2 show exact agreement agree for candidates 4+9*
- *Rater 3 is more severe than rater 1+2*
- *Rater 3 has no exact agreements with raters 1+2.*
- *The rater trends between raters 1-3 are very similar (infit)*



BRITISH COUNCIL

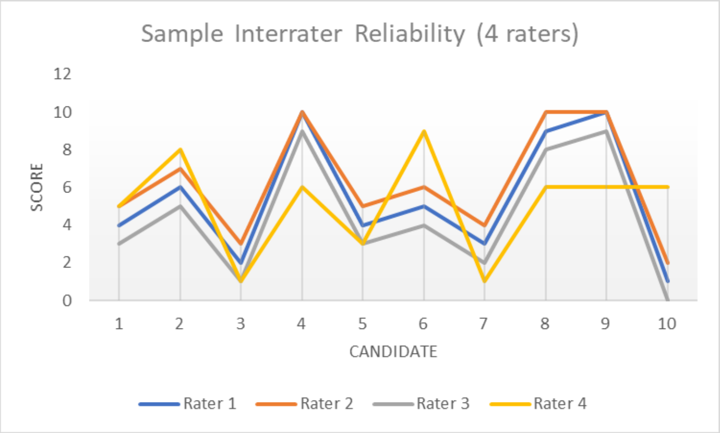


East Asia Assessment
Solutions Team

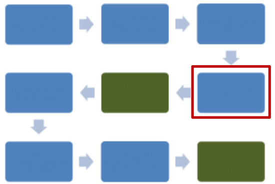
Rater Reliability

- FACETS software used to calculate the degree of agreement between rater scores (interrater reliability)
- Three main reliability factors were considered, **separation** (leniency severity) **fit,** and **exact agreements**
- These three principles can be illustrated using the following (fictional) example:

- *Rater 1 is generally more severe than rater 2*
- *Rater 1 and 2 show exact agreement agree for candidates 4+9*
- *Rater 3 is more severe than rater 1+2*
- *Rater 3 has no exact agreements with raters 1+2.*
- *The rater trends between raters 1-3 are very similar (infit)*
- *Rater 4 has poor fit with raters 1-3. In this study we would consider retraining or removing rater 4 data.*



East Asia Assessment
Solutions Team



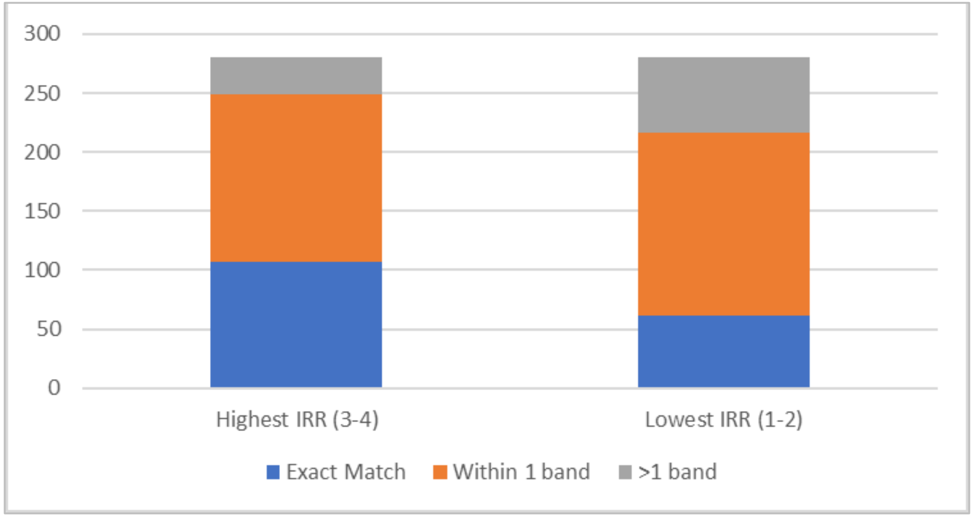
Rater Reliability

Rater Measurement Statistics

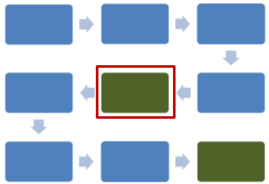
Rater	Fair Average	Infit (MnSQ)
1	3.27	0.79
2	4.15	1.01
3	3.7	1.1
4	3.41	0.69
5	3.79	1.45
S.D. (5 raters)		.63

Agreement (exact matches) 5 Raters	
Opportunities	2800
Expected	1036 (37%)
Observed	926 (33.1%)

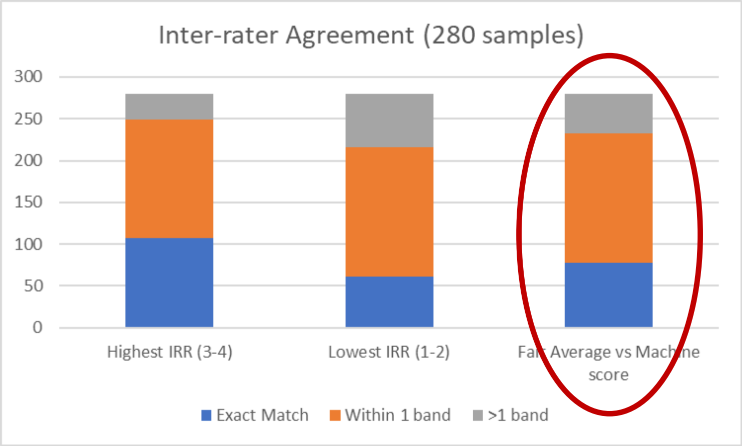
Rater Agreement (highest/lowest) 280 samples

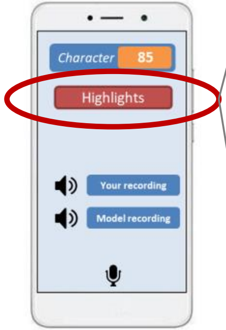


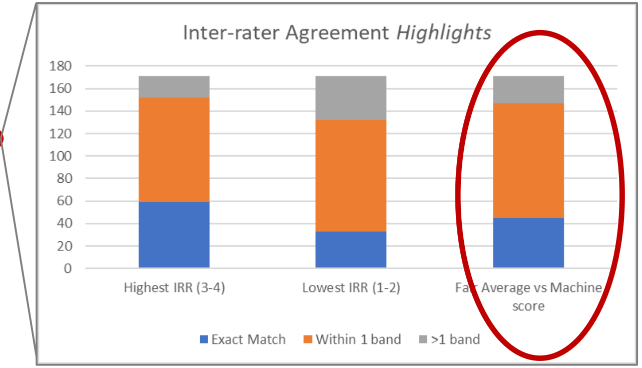
Human-machine Agreement

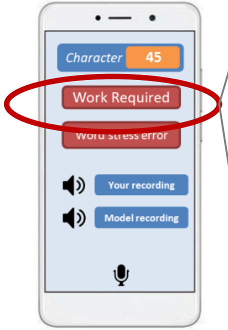


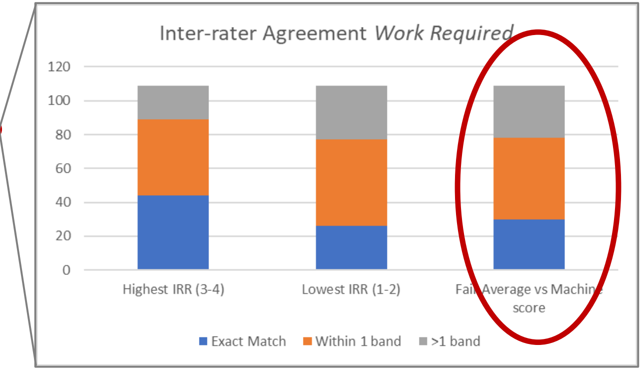
East Asia Assessment
Solutions Team



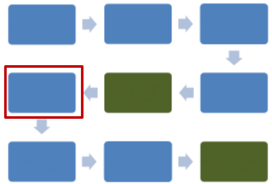








East Asia Assessment
Solutions Team



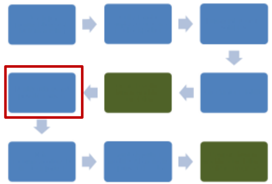
Identifying Outliers: Word Level

- The human and machine ratings per word were ranked from highest to lowest and the correlation calculated.
- Word groups with a negative correlation between human and machine ratings were collated and sent to the tech partner for further investigation.

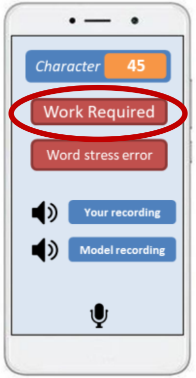
Word	Syllables	Sample (no. of recordings)	Correlation	Rank Av diff
Literature	4.00	13	-0.170	98
Talented	3.00	7	-0.086	76
Reduces	3.00	6	-0.619	111
Novel	2.00	4	-0.146	77
Page-turner	3.00	4	-0.830	104
Character	3.00	3	-0.932	210

East Asia Assessment
Solutions Team

East Asia Assessment Solutions Team



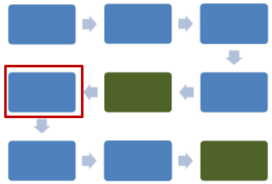
Focus on Work Required



False Positive

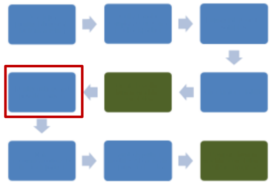
False Negative

East Asia Assessment Solutions Team



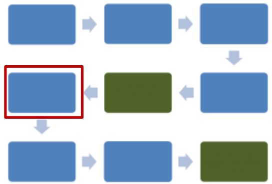
Identifying Outliers: Response Level

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



Identifying Outliers: Response Level

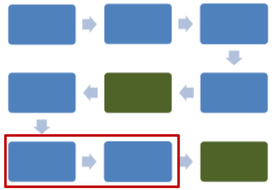
Word	Comments	Type	Rater Fair Average	Machine Score	Error filter	Feedback Error
Visit	Machine score significantly lower than human average. A review suggests a native speaker controlling stress and phonemes correct as per UK RP.	Work required	4.55	3.19	Phoneme	False negative
character	Machine score significantly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



Identifying Outliers: Response Level

Word	Comments	Performance	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. Machine underrating native/near native performance.	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. Conclusion: need to incorporate error filter into the overall pron score.	Work required	3.40	4.96	Stress	False positive
easy-going	Machine score significantly higher than human average. A review shows word stress is poorly managed. This finding is supported by the Machine syllable stress error filter result. Conclusion: need to incorporate error filter into the overall pron score.	Work required	3.40	4.98	Stress	False positive
page-turner	Machine scores significantly lower than human average. A review suggests a native speaker controlling stress and phonemes correct as per UK RP. Machine underrating native/near native performance.	Work required	4.74	3.64	Phoneme	False negative
fiction	Machine scores significantly lower than human average. A review suggests a native speaker controlling stress and phonemes correct as per UK RP. Machine underrating native/near native performance.	Work required	4.83	3.80	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. Conclusion: need to incorporate error filter into the overall pron score.	Work required	3.82	4.98	Stress	False positive

East Asia Assessment Solutions Team



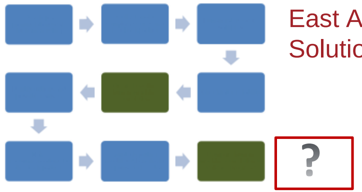
Recommendations and modifications

Recommendations (test developers):

1. Check the engine performance of the following outlier words:
   (a) Literature (b) Talented (c) Reduces
   (d) Novel (e) Page-turner (f) Character
2. Check engine performance in relation to standard UK variant performance
3. Incorporate the error filter (phoneme/word stress) into the overall scoring model

Modifications (tech partner):

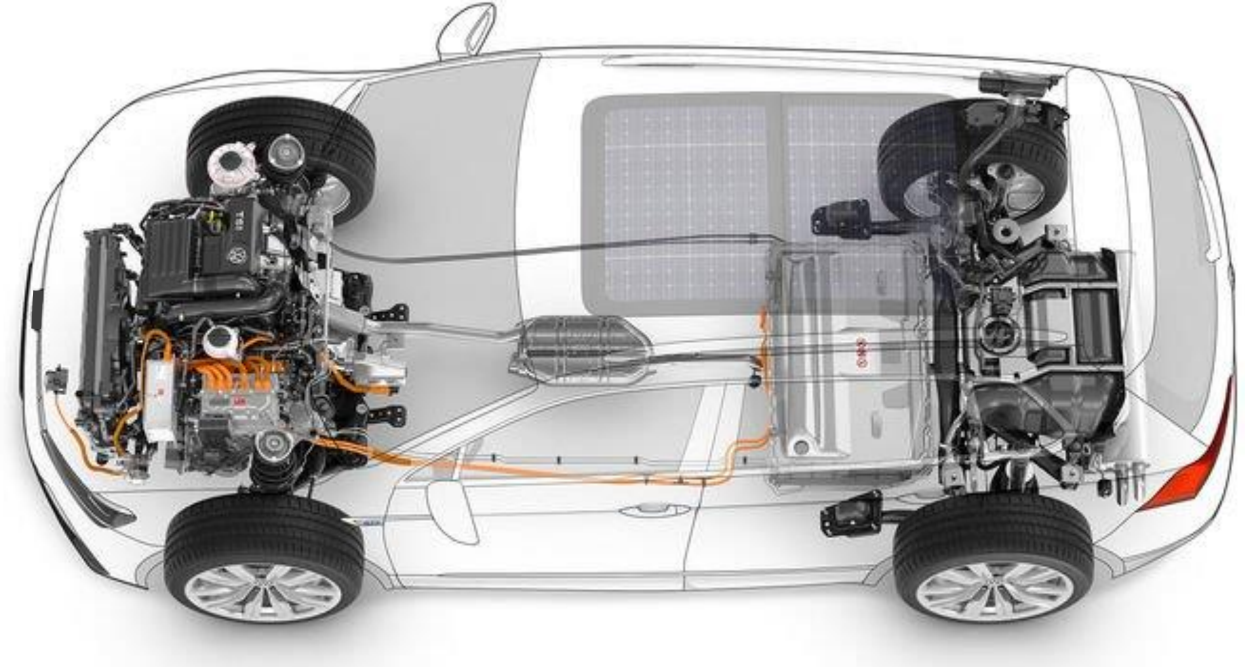
1. Dictionary entries for the outlier word list reviewed, corrected phoneme error in entry “literature”
2. Upgraded dictionary entries to ensure UK English variant was included
3. Error detection filter was combined with the overall scoring mechanism. Penalty score was deducted from the overall score for each phoneme/stress error made



Next Steps

1. Test the new combined model with a wider range of vocabulary items and learner samples
2. Validate penalty mechanism and work required/highlights categories
3. Extend study to include longer utterances and broader construct of pronunciation (liaison, intonation, sentence stress)
4. Conduct a more targeted validation study aligning human and machine pronunciation features and scales (comparing like-with-like)

What's under the bonnet?



http://www.caradvice.com.au

BRITISH COUNCIL

East Asia Assessment Solutions Team

Thank You!

Any Questions?

trevorjohn.breakspear@britishcouncil.org.cn

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

BRITISH COUNCIL

Rubric features

- Rating scale design included 4 categories; (1) task engagement; (2) fluency; (3) lexicogrammar; (4) pronunciation
- A set group of 6 raters was divided into two groups to score the test; each sample was scored by at least three raters
- A selection of anchor items were scored by all 6 raters to aid facets analysis and data optimisation

CEFR	Bands
B2	6
	5
	4
	3
	2
	1
A0	0

Band	Descriptor
4	*Speaker is able to respond relevantly to all aspects of the prompts in a generally clear, coherent manner* *The speaker will typically demonstrate:* *-Steady delivery with some hesitation and searching for words* *-Range of simple vocabulary with some successful paraphrasing* *-Mix of simple and complex forms with noticeable errors* *Speaker is intelligible, with occasional lapses*

East Asia Assessment
Solutions Team

Autorater readiness for different speaking tasks

Task	Functions (national curriculum+ CSE)	Feature domains	AI Readiness*
Read aloud	Can speak with relatively accurate pronunciation and appropriate intonation	Phonology (PN)	Mature
Question and answer	Can communicate on familiar topics using simple language.	PN/vocabulary and grammar (V+G)	Developing
Structured narrative	Can provide information about personal backgrounds and experiences. Can tell simple and short stories.	PN/V+G/discourse/ task achievement	Developing
Conversation based (virtual interlocutors)	Can take part in simple role plays with the help of the teacher.	PN/V+G/interaction/ discourse/task achievement	Initial development

*Adapted from Evanini, 2017

Autorater readiness for different speaking constructs

Speaking features	Functions
• Phonology	• Can speak with accurate pronunciation and appropriate intonation in the above tasks.
• Vocabulary and grammar	• Can communicate on familiar topics using simple language. • Can use simple language to describe experiences with support. • Can take part in role plays with support.
• Interaction • Discourse • Task achievement	• Can introduce the topic and maintain the conversation with several turn-takings.

Mature	Developing	Initial Development

(Breakspear, Lam, Khabbazbashi, Chan, 2018)