BRITISH COUNCIL

BAAL TEASIG Online Conference

Validating Auto-Scoring Models for Use in Language Assessment

Barry O'Sullivan, Trevor Breakspear & William Bayliss – September 17th 2021

Overview

Context: ISL placement test – Auto-scored – Low Stakes – China

Objective 1: To learn about the specifications of the scoring system

Objective 2: To validate the system

This Paper: Focus on one element of the project – Model Card Production

ISL Background
Model Cards
Completing the Card

IELTS Smart Learning (ISL) (BOS)

Concept:
IELTS Speaking test preparation App
Structured/graded programme
Multiple staged learning units [we are focusing on the placement element]
Auto-Scored
Formative feedback

Ownership:
IELTS Partnership
Developed by the partners from a British Council PoC

Target Users:
Chinese learners preparing to take IELTS

Availability:
China only (currently)

Model Cards Overview

Proposed Margaret Mitchel et al. (2019)

Designed for use for AI driven or supported high-stakes decisions

Designed to identify the types of evidence required for model use validation

Validation seen as aimed at a largely non-technical audience

Model Cards Details

Element
Model Details
Intended Use
Factors
Metrics
Training Data
Evaluation Data
Quantitative Analyses
Ethical Considerations

Model Cards Details

Element	Description
Model Details	basic information about the model
Intended Use	specific use cases envisioned during development
Factors	potential sources of bias
Metrics	measures chosen to reflect potential real-world impacts of the model
Training Data	dataset(s) used to build the system
Evaluation Data	dataset(s) used to evaluate the system
Quantitative Analyses	evidence to support the use of the system
Ethical Considerations	

Model Cards Details: Intended use

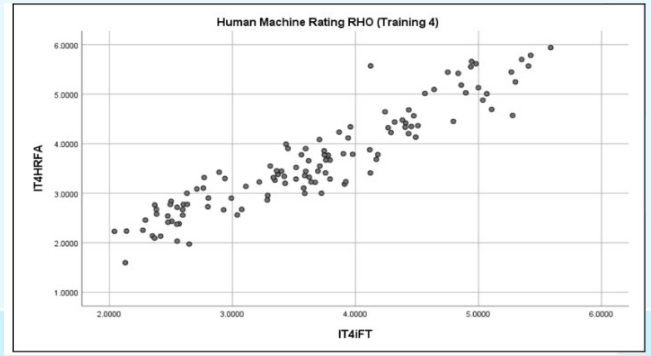
Element	Description
Model Details	basic information about the model
Intended Use	A **low-stakes** placement test recommending starting practice level. Recommendation can be ignored by the user. • L1 Chinese users • 14-19 years • All regions of China
Not intended	• High-stakes decision on performance • Non-Chinese L1 users • Young learners (<13)
Factors	potential sources of bias
Metrics	measures chosen to reflect potential real-world impacts of the model
Training Data	dataset(s) used to build the system
Evaluation Data	dataset(s) used to evaluate the system
Quantitative Analyses	evidence to support the use of the system

Model Cards Details: factors

Element		Description
Model Details		basic information about the model
Intended Use		specific use cases envisioned during development
(Risk) Factors	Population	**Subgroups:** gender/ age /city tier/user L1/ accent **Environment:** multiple user voices; quality of the hardware, connection and bandwidth.
	Data	**Metadata coverage:** 52% gender/ 34% age/ 39% included city **Findings in relation to intended use:** • Training data older than the target demographic (16-19). • Strong representation in tier 1 and 2 cities.
Metrics		measures chosen to reflect potential real-world impacts of the model
Training Data		dataset(s) used to build the system
Evaluation Data		dataset(s) used to evaluate the system

Model Cards Details: metrics

Element	Description
Model Details	basic information about the model
Intended Use	specific use cases envisioned during development
Factors	potential sources of bias
Metrics	• Human Inter-rater reliability (IRR): ✓ Within 1 band = **91.6%** • Human-Machine reliability (HMR): ✓ Spearman's r_s = .93 (p= <.001) ✓ Within 1 band = **100%**
Training Data	dataset(s) used to build the system
Evaluation Data	dataset(s) used to evaluate the system
Quantitative Analyses	evidence to support the use of the system
Ethical Considerations	



Model Cards Details: quantitative analyses

<table>
<tr><th>Element</th><th>Description</th></tr>
<tr><td>Model Details</td><td>basic information about the model</td></tr>
<tr><td>Evaluation Data</td><td>dataset(s) used to evaluate the system</td></tr>
<tr><td>Quantitative Analyses</td><td>Human rating data
<table>
<tr><th>Source</th><th>Mean Square</th><th>F</th><th>Sig.</th></tr>
<tr><td>Age * Tier</td><td>.224</td><td>.314</td><td>.815</td></tr>
<tr><td>Gender * Tier</td><td>.189</td><td>.265</td><td>.607</td></tr>
<tr><td>Age * Gender * Tier</td><td>.140</td><td>.196</td><td>.659</td></tr>
</table>
Machine rating data
<table>
<tr><th>Source</th><th>Mean Square</th><th>F</th><th>Sig.</th></tr>
<tr><td>Age * Tier</td><td>.053</td><td>.089</td><td>.966</td></tr>
<tr><td>Gender * Tier</td><td>1.759</td><td>2.923</td><td>.090</td></tr>
<tr><td>Age * Gender * Tier</td><td>.123</td><td>.205</td><td>.652</td></tr>
</table></td></tr>
</table>

Model Cards Details: Intended use

Element	Description
Model Details	basic information about the model
Intended Use	A **low-stakes** placement test recommending starting practice level. • L1 Chinese users • 14-19 years • All regions of China
Not intended	• High-stakes decision on performance • Non-Chinese L1 users • Young learners (<13)
Factors	potential sources of bias
Metrics	measures chosen to reflect potential real-world impacts of the model
Training Data	dataset(s) used to build the system
Evaluation Data	dataset(s) used to evaluate the system
Quantitative Analyses	evidence to support the use of the system

Lessons Learnt & Conclusions

Lessons Learnt

Don’t try this at home…

This work can only be done in partnership

Significant learning is required from all partners

The process can help us to improve our scoring systems

The Model Card approach helps us to know what questions to ask

and how to ask them!

Conclusions

We urgently need to understand more about the AI-scoring systems currently in use

We also need to consider how AI-scoring systems fit into our validation models and processes

Model Cards offers a realistic approach to model validation representation and should be broadly adopted across language testing

The process offers test developers and AI-developers an opportunity to develop a common language and common expectations on AI model use

Thank You

BRITISH COUNCIL

ISL Model Card

Model Details

This relates to the basic information about the model. For example, information about:

- the developer(s)
- date of development
- the type of model it is (Long Short-Term Memory (LSTM), linear regression model etc.)
- model training details
- research or descriptive documentation (academic papers; Kaggle resources; vendor specific docs.)
- citation and license details
- contact details

Model Details ISL

Developers	iFLYTEK Research, China
Date of Development	Model design developed in 2017 ISL version developed between July 2019 and February 2020.
Type of Model	Long Short-Term Memory (LSTM), linear regression model. An automatic speech recognition (ASR) sub-model converts users' speech to text format. This output is then converted into vectors using LSTM. Finally, multi regression models such as linear regression and decision tree regression are trained by human raters' scores. The model predicts the L2 English speaking performance (score) from a series of prompts—based on the holistic evaluation of features including task engagement, fluency, lexical resource, grammar and pronunciation.
Model Training	Model(s) trained by IFLYTEK (human rating data optimised for machine learning by the British Council).
Research	The raw untrained model —iteration one— underwent four subsequent sessions of training. The current model is its fifth iteration.

Intended Use

Here the specific use/user cases envisioned during development are clearly described.

Also included here will be any use cases that are out of scope.

Intended Use ISL

Intended	A low-stakes formative measure to streamline IELTS Smart Learning (ISL) users (aged 14-19 & based in China) to the most appropriate IELTS materials and practice tasks based on their ability. The recommendation is not compulsory; the user can choose to ignore it and start from any level in the product.
Not Intended	High-stakes decisions about learners

Factors

Potential sources of bias

Should (where appropriate) include variables related to:

- demographics
- phenotype group (e.g. height, weight, colour)
- environmental conditions (e.g. with or without background interference),
- technical attributes

Factors ISL

Risk Factors – Population	These include groups for gender, age and city tier; uniqueness of the users' answering – multiple user voices; proficiency outliers; users' L1 or dialect background, and relevant ASR accuracy; user speech issues such as stammers; users' recording environment; the quality of the hardware, connection and bandwidth.
Risk Factors – Data	Total training + evaluation dataset samples n=550 52% of samples included gender metadata (32% female, 20% male, 48% unknown), 34% of samples included age metadata. Most training data comes from learners older than the target demographic (16-19). Samples with School/Institution city tier metadata accounts for 39% of the total samples, with strong representation among tier 1 and 2 cities

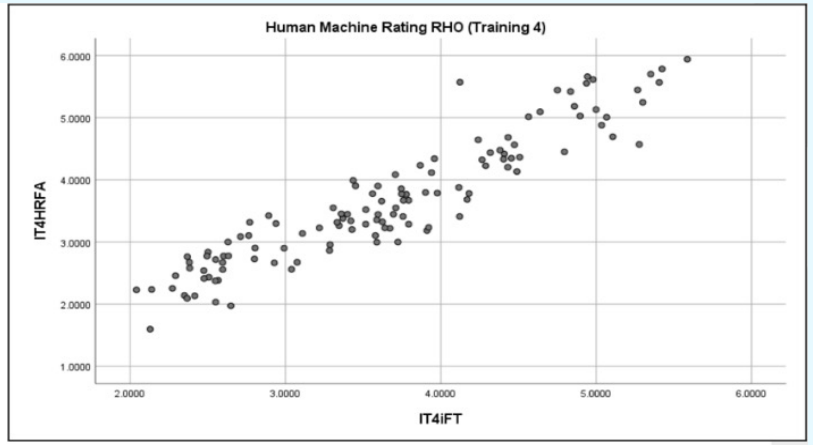
Metrics

These should “reflect real world impacts of the model’ (Mitchell et al., 2019: 3)

May also include:

- model performance metrics
- decision thresholds
- variation approaches.

Metrics ISL

Model Performance	Evaluation metrics include: • Human Inter-rater reliability (IRR): ○ Exact – 42.5% ○ Up to1 Band – 91.6% ○ >1 Band – 8.4% • Human-Machine reliability (HMR): ○ RHO – 0.93 ○ Ave H-M difference – 0.25 ○ Exact Agreement – 67% ○ Up to 1 Band – 100% 
Decision Thresholds	Above outcomes reflect decision thresholds (placement use)
Variation Approaches	None applied

Evaluation Data

Overviews the datasets that are used in the quantitative analysis presented in the card, and includes:

- information on the datasets
- the motivation behind the decisions to select these datasets
- any data cleaning and processing used.

Training Data

Where feasible, this should be presented and should mirror the variables recorded in the evaluation data section – though here presents specific numbers as opposed to general descriptions.

Evaluation & Training Data

Initial Model Training Data	Good coverage of the ISL user demographic features (ethnicity/age). Over 50 million English recordings were used to build and optimise the model. Several million recordings per year are used to continually optimise accent representation within the model.
ISL Model Training Data	IFLYTEK collected a total of 923 recordings, of which 570 were human rated. Post data optimisation, 433 for model training.
Evaluation Data	117 recordings used for model evaluation. Iterative approach through 4 generations saw proficiency representation controlled to avoid focus on the most common scores (A2/B1). This improved human-machine reliability across the scale range and particularly at the tails.

Quantitative Analysis

This should include results relating to:

- unitary (e.g. gender bias or age bias)
- intersectional (e.g. gender by age bias)

Quantitative Analysis ISL

Overview	A series of ANOVAs were used to explore the different independent variables (one way), and the interaction between the independent variables (3 way). Statistically significance may suggest the presence of bias in the human or machine rating
Unitary	The one-way ANOVAs results suggest there is no statistically significant difference between the means of the different levels of the ‘age buckets’, ‘city tier’ or ‘gender’ in either the human or machine scores. The only variable that is close to statistical significance is city tier for human ratings. **No evidence of bias by subgroup based on this dataset.**
Intersectional	The results of the three-way ANOVA suggest that there is no statistically significant effect for the interaction the variables studied — ‘age buckets’, ‘city tier’ and ‘gender’. **No evidence of bias by subgroup based on this dataset.**

Quantitative Analysis ISL (ANOVA Human Output)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6.453[a]	13	.496	.697	.763
Intercept	402.836	1	402.836	565.474	.000
Age	.978	3	.326	.458	.712
Gender	7.791E-6	1	7.791E-6	.000	.997
Tier	.624	1	.624	.876	.351
Age * Gender	2.693	3	.898	1.260	.292
Age * Tier	.672	3	.224	.314	.815
Gender * Tier	.189	1	.189	.265	.607
Age * Gender * Tier	.140	1	.140	.196	.659
Error	79.787	112	.712		
Total	1434.199	126			
Corrected Total	86.240	125			

a. R Squared = .075 (Adjusted R Squared = -.033)

Quantitative Analysis ISL (ANOVA Machine Output)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	8.026[a]	13	.617	1.026	.432
Intercept	423.169	1	423.169	703.383	.000
Age	1.825	3	.608	1.011	.391
Gender	.214	1	.214	.355	.552
Tier	1.228	1	1.228	2.042	.156
Age * Gender	2.850	3	.950	1.579	.198
Age * Tier	.160	3	.053	.089	.966
Gender * Tier	1.759	1	1.759	2.923	.090
Age * Gender * Tier	.123	1	.123	.205	.652
Error	67.381	112	.602		
Total	1475.660	126			
Corrected Total	75.407	125			

a. R Squared = .106 (Adjusted R Squared = .003)

Ethical Considerations

Reflects ethical issues current to the context and will include (but not exclusively):

- sensitive data usage
- high-stakes decision usage
- risk and potential negative impact concerns
- potentially toxic use cases.

Ethical Considerations ISL

Sensitive Data	Recorded speech data is considered personal data under GDPR; Some under 18 samples were used in the training data; however, all users must sign a robust privacy agreement before using the App.
Data Cleaning	Model iteration 4 was used to filter out median performances, to ensure more representation of high and low proficiency performances.
Unused Data	Anomalous recordings —corrupt files, excessive background noise — filtered prior to human rating.

Caveats & Recommendations

This section should list any additional issues not covered in the previous sections – typically emerging from the analyses that contribute to the training of the model.

Caveats & Recommendations ISL

Data Quality	Ideal evaluation and training datasets to include age, tier and gender metadata for all data points.
Unsuitable Users	Not suitable for learners significantly different from those represented in this development.
Unsuitable Task Use	Not suitable for rating language beyond the scope of the model specific prompts.
Unsuitable Decision Level	Model is designed for use as a low-stakes learning aid to stream students to a recommended start point in the app content. It should not in its current form be used for high-stakes language assessment.
Possible expansion of use	With further speech data and robust benchmarking studies linking reported scores with the CEFR the model may be used for mid or high stakes assessment.