

# Supporting the Regulator...

*What about the rating?*

Our perspectives on regulators & rating issues

A case study in cross-rating  
between 2 UK CAA-approved TSPs

Summary of helpful considerations for regulators



**MAYFLOWER**  
**COLLEGE**



**TEST OF ENGLISH  
FOR AVIATION**

Operating in 4 continents  
Approved in many States

...including the UK CAA



30,000+ licensing tests

**Does Chapter 6 of 9835**

**‘Language Testing Criteria for Global Harmonization’**

***truly* support the regulator..?**

6.3.4.1: remote or live rating ok ✓

6.3.4.2: better to have 2 raters ✓

6.3.4.3: important to assess rater reliability ✓

6.3.4.4: speech recognition technology ok ✓

If ***you*** were accountable for approving tests,  
what questions would you be asking about  
a TSP's approach to rating ?

- *What is a 'L4 performance' on your test?*
  - *What do you do to check rating reliability?*
  - *...and to improve reliability?*
- 
- How *open* is the TSP about their rating?
  - Is L4 with TSP~~X~~ generally also a L4 with TSP~~Y~~?

# Rating Standardisation Pilot Project





## Project Objectives

- assess level of rater agreement between 2 active CAA-approved TSPs
- further understanding of fellow TSP work
- activate further work on performance descriptions (internal & external, where necessary)
- assess possibility of larger project to include all CAA-approved TSPs

## Project Design

Each TSP provided:

- 5 full, anonymised tests of UK-licensed candidates (labelled *Candidate 1*, *Candidate 2*, etc.)
- 5 sets of original scores (labelled *Set A*, *Set B*, etc.) for each performance
- Full description of test's assessment criteria

## Pre-Project

Each TSP:

- Signed project agreement
- Signed confidentiality agreements
- Agreed to respect integrity of both tests & adhere to ILTA Code of Ethics
- Transferred materials by secure server

## Task Design











Each TSP's Senior Rating Team agreed to:

1. Study & discuss assessment criteria
2. Rate 5 tests (discuss & agree 6 profile scores for each performance)
3. Compare to *Score Sets* & discuss completion of table before submission to TSP partner for analysis...



# Scores

# Rating Project

Candidate	Original scores (P S V F C I)						Scores from other TSP (P S V F C I)					
 Anglo-Continental <b>A</b>	3	4	4	4	3	4	4	4	4	4	3	4
 Anglo-Continental <b>B</b>	5	5	5	5	5	5	6	5	5	6	5	6
 Anglo-Continental <b>C</b>	4	4	4	4	5	5	5	5	5	5	5	5
 Anglo-Continental <b>D</b>	5	5	5	5	4	5	5	5	5	5	4	5
 Anglo-Continental <b>E</b>	4	5	4	5	4	5	5	4	4	4	4	4
 MAYFLOWER COLLEGE <b>A</b>	6	5	5	5	6	6	6	6	6	6	6	6
 MAYFLOWER COLLEGE <b>B</b>	4	3	3	3	3	3	3	3	3	3	3	3
 MAYFLOWER COLLEGE <b>C</b>	5	5	5	4	5	5	4	5	5	4	5	4
 MAYFLOWER COLLEGE <b>D</b>	4	4	4	4	4	4	4	4	4	4	4	4
 MAYFLOWER COLLEGE <b>E</b>	5	5	6	5	4	6	5	5	5	5	5	6

## Results

- **Anglo Continental's team** correctly matched 5 performances to score sets
- **Mayflower College's team** correctly matched 3 performances
- 3 ICAO Overall Score disagreements – only 1 considered '*unreasonable*' rating
- Correlations for rating of 3 profiles high

## Means

sample size = 10 tests

	 Anglo-Continental	 MAYFLOWER COLLEGE
<b>Pronunciation</b>	<b>4.30</b>	<b>4.90</b>
<b>Structure</b>	<b>4.60</b>	<b>4.50</b>
<b>Vocabulary</b>	<b>4.50</b>	<b>4.60</b>
<b>Fluency</b>	<b>4.50</b>	<b>4.50</b>
<b>Comprehension</b>	<b>4.40</b>	<b>4.30</b>
<b>Interactions</b>	<b>4.70</b>	<b>4.80</b>
<b>ICAO Overall</b>	<b>4.20</b>	<b>4.10</b>



## Correlations (Pearson)



	P	S	V	F	C	I	ICAO
P	.83						
S		.75					
V			.78				
F				.70			
C					.95		
I						.84	
ICAO							.79



## Disagreements



### Candidate C

L4+/L5 borderline decision **P S V & F**



### Candidate A

L5+/L6 borderline decision on **S V & F**

## *Unreasonable* Rating



MAYFLOWER  
COLLEGE

**Candidate E:** awarded L4 for C ...

*Anglo-Continental team felt  
assessment itself fair but  
Comprehension assessment criteria  
may be unreasonably harsh...*

## Difficulties & Constraints

- 10 tests = small sample for meaningful data analysis
- Matching task means 1 incorrect match = 2 incorrect
- Difficulties in rating partner tests without guidance

## Project Outcomes

- Professionally meaningful & awareness-raising
- Intra-TSP review on descriptions of typical level indicators (esp. levels 5 & 6) would be beneficial
- Further inter-TSP work on **S**, **V** & **F** rating beneficial
- CAA-led standardisation project desirable



# Action

- Greater awareness through open collaboration
- Reviewing & Re-writing internal performance descriptions
- Conducting research with all **TEA** Examiners into Comprehension assessment method
- Pushing for more CAA-approved collaborations

# Summary: *What can regulators do?*

- Host meetings of approved TSPs / encourage open collaboration (& discourage 'commercialisation' as far as possible)
- Support inter-TSP standardisation
- Observe tests
- Conduct random test sampling
- Ask for detailed descriptions of candidate performance indicators
- Show interest in the rating process!



MAYFLOWER  
COLLEGE

Please say *Hi* to me or our testing partners here at the workshop







MAYFLOWER  
COLLEGE

***Many thanks***

**ben@maycoll.co.uk**



Extra slides...

# Proposal for larger CAA Standardisation Project

- all CAA-approved TSPs invited to simplified project
- objective: external standardisation leading to *internal* outcomes
- each TSP provides 3 tests, original scores & assessment criteria
- each TSP Rater Team assesses scores as '*unreasonable*' or '*not unreasonable*' (with additional comments)
- no large data analysis
- results for internal use only