

Supporting the Regulator...

What about the rating?



Operating in 4 continents
Approved in many States...

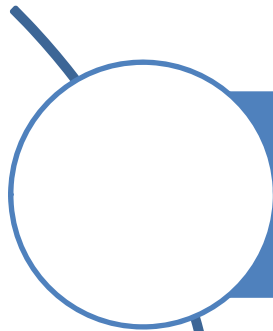
...including the UK CAA



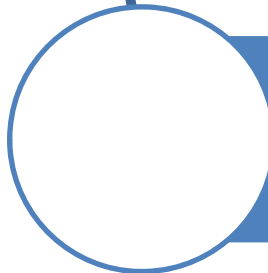
30,000+ licensing tests



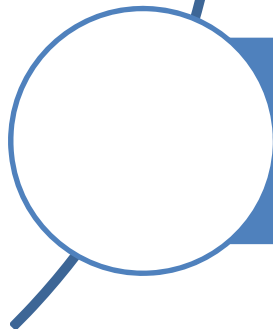
MAYFLOWER
COLLEGE



Our perspectives on regulators & rating issues



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



Summary of helpful considerations
for regulators

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



Rating Project

assess level of rater agreement between 2 CAA-approved TSPs

further understanding of fellow TSP work

Objectives

activate further work on performance descriptions

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*
4. Discuss completion of task table before submission to TSP partner for analysis...











Rating Project

Candidate	Matching Score Set (A – E)?	Original scores (from that Score Set)						In the Senior Rater Team’s opinion, is the scoring <i>unreasonable</i> , or <i>not unreasonable</i> ?	Agreed scores from the Senior Rater Team						Additional Comments
		P	S	V	F	C	I		P	S	V	F	C	I	
1		P	S	V	F	C	I		P	S	V	F	C	I	
2		P	S	V	F	C	I		P	S	V	F	C	I	
3		P	S	V	F	C	I		P	S	V	F	C	I	
4		P	S	V	F	C	I		P	S	V	F	C	I	
5		P	S	V	F	C	I		P	S	V	F	C	I	

‘unreasonable’ or ‘not unreasonable’ rating

Scores

Rating Project

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

Results Summary

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

Rating Project

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...*

Rating Project

Means

sample size = 10 tests

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

Correlations (Pearson)

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

Difficulties & Constraints

- 10 tests = small sample for meaningful data analysis
- Matching task: 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!*



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



BULATSA
BULGARIAN AIR TRAFFIC SERVICES AUTHORITY



Many thanks

شكرا جزىلا

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