EXPLORING RELIABILITY IN ASSESSING LANGUAGE EXAM PERFORMANCE IN THE GOETHE-INSTITUT'S WORLDWIDE NETWORK OF RATERS

PARIS, APRIL 2014

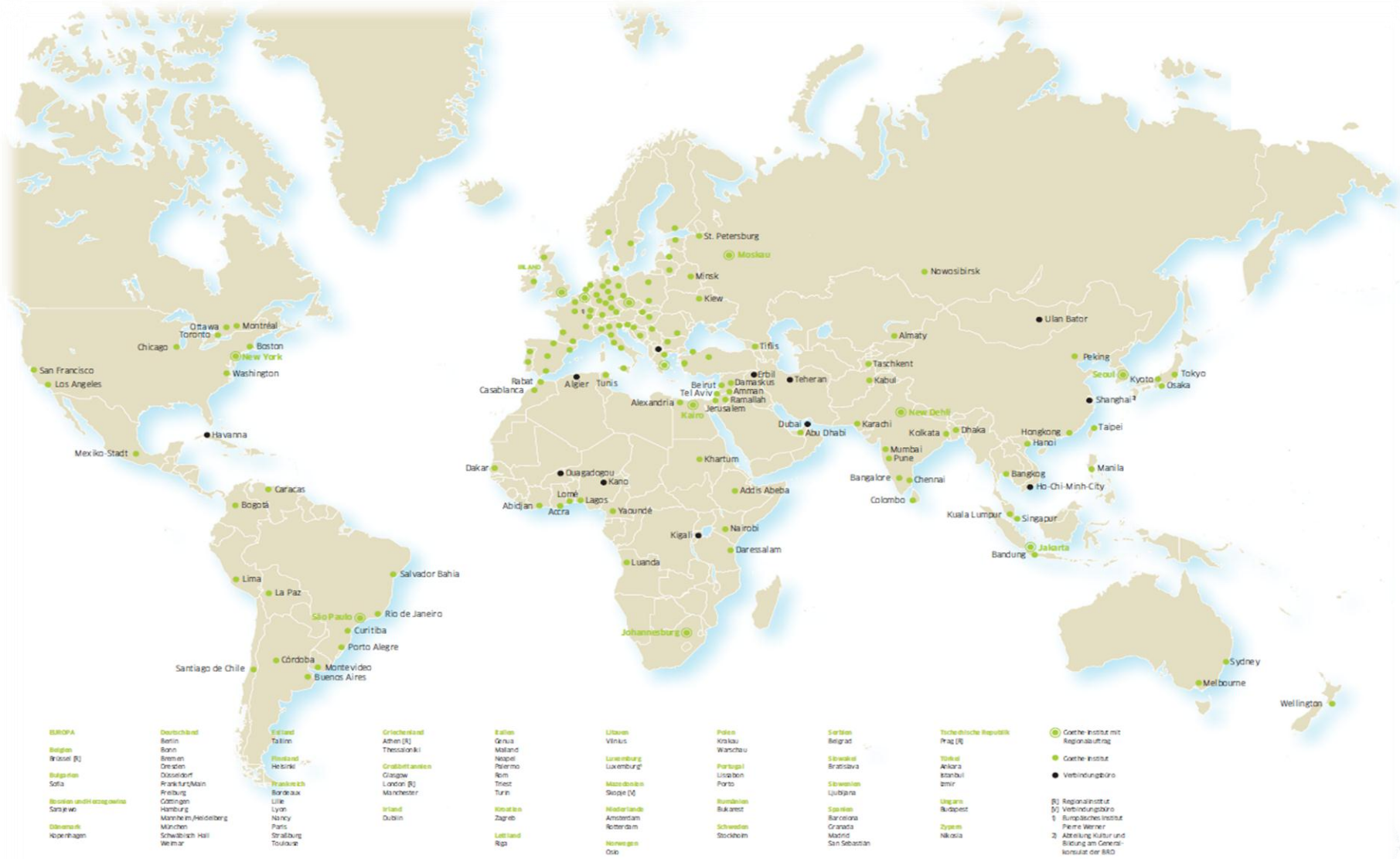
CHRISTINA GREGOR GOETHE INSTITUT
JANE LLOYD ALTE VALIDATION UNIT

GOETHE INSTITUT
Sprache. Kultur. Deutschland.

AGENDA

1. **STANDARDISATION OF RATING**
    1. The Goethe-Institut's network
    2. ALTE minimum standards
    3. System of training and monitoring
    4. Certification of raters
    5. Live-test Analysis
2. **ANALYZING DATA**
    1. Certification of raters
    2. Live-test Analysis
3. **NEXT STEPS?**

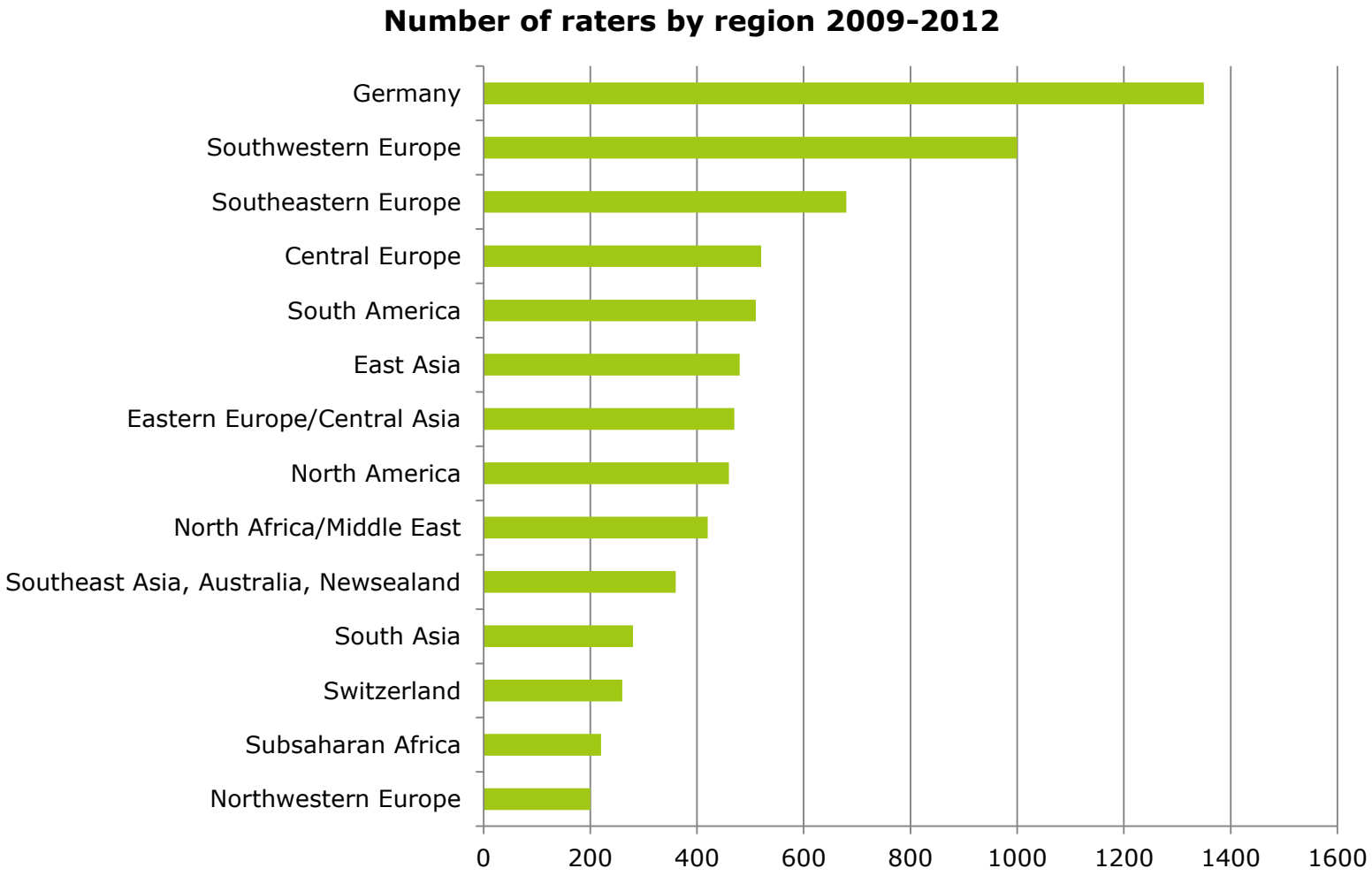
THE GOETHE-INSTITUT'S NETWORK



THE GOETHE-INSTITUT'S NETWORK

- 500 testing locations in 93 countries: 150 Goethe-Instituts, 350 examination partners
- 2013: 246,000 German language examinations (A1-C2) and rising
- 7,000 raters worldwide assess language exam performance locally

THE GOETHE-INSTITUT'S NETWORK



ALTE MINIMUM STANDARDS

Relevant minimum standards:

11

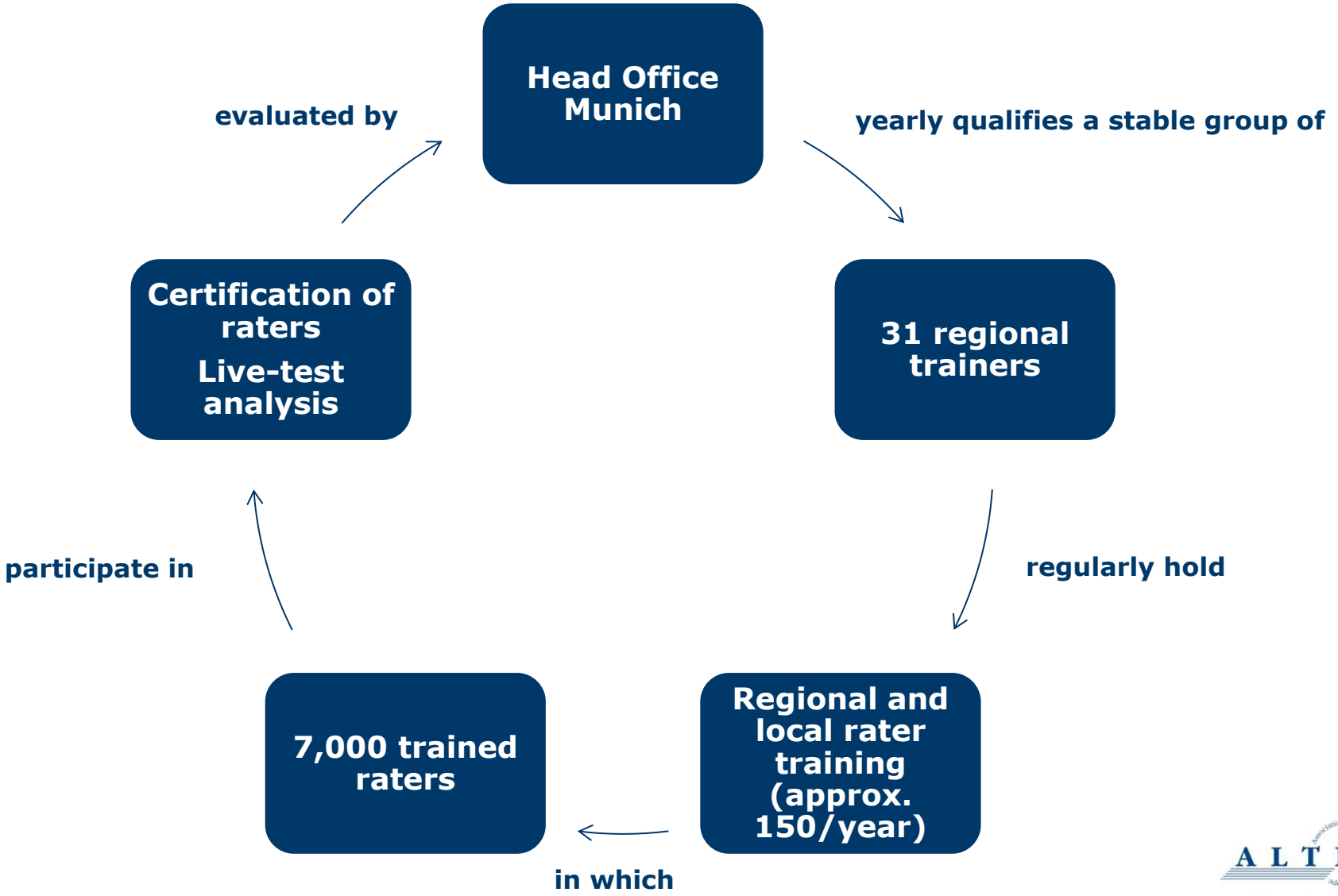
- Marking is sufficiently accurate and reliable for purpose and type of examination.

12

- You can document and explain how marking is carried out and reliability estimated, and how data regarding achievement of raters of writing and speaking performances is collected and analysed.

➔ How to guarantee reliability in assessing the German examinations of the Goethe-Institut?

SYSTEM OF TRAINING AND MONITORING



CERTIFICATION OF RATERS („PRÜFERZERTIFIKAT")

OBJECTIVE

Proof of standardisation of raters
Per examination (A1-C2): 3 written samples, 1 video of an oral exam

METHOD

Data collection: each rater in the GI-network has to participate via platform „Moodle"

Setting the marks: team of experts in the head office rates the samples

Data Analysis

RESULT

No significant deviation from the standard: certificate valid for 5 years

Significant deviation from the standard: retraining and recertification

EXAMPLE SOUTHWESTERN EUROPE, GI PARIS

GI Paris				
Rater 998	2009	2010	2011	2012
A1 (F1) W	!	✓		
A2 (F2) W	✓			
A1 (SD1)O	✓			
A1 (SD1) W	✓			
A2 (SD2) O	✓			
A2 (SD2) W	✓			
B1 (ZD) O	✓			
B1 (ZD) W	✓			
B2 O	!	!		✓
B2 W	!	✓		✓
C1 O	✓			
C1 W	✓			
C2 O	✓			
C2 W	✓			

ANALYZING RATERS

1. Rating based on sample materials: certification

- Comparison with the Goethe Institut standard
- Acceptable level of correlation to the standard
- Many Facet Rasch Measurement
- Severity Measure

2. Live test ratings

- Reliability of scores assigned by pairs of raters
- Analysis of inter-rater correlation

ANALYZING RATERS

1. Rating based on sample materials: certification

- Comparison with the Goethe Institut standard
- Acceptable level of correlation to the standard
- Many Facet Rasch Measurement
- Severity Measure
- Cut off based on standard deviation from the rating
- Measurement Error taken into account

2. Input

- Excel Spreadsheet with rater, centre, exam, criteria, grade

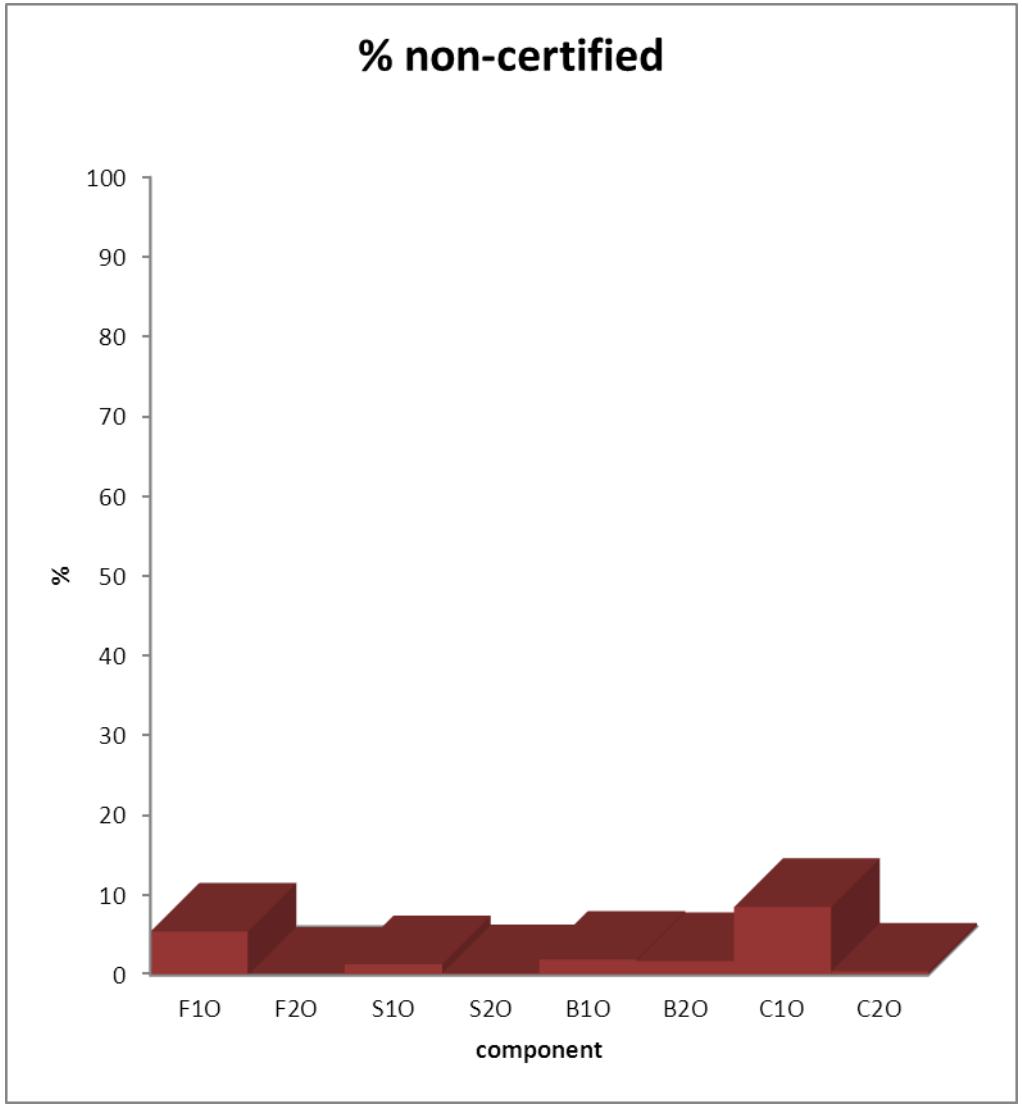
3. Outputs

- Overviews arranged by exam, skill, centre and examiner

RESULTS 2013: SPEAKING

exam	cut offs low	high	lenient	severe	non-certified count	total analysed	% non-certified
F1o	-1.645	1.645	1	39	**40**	732	5.464
F2o	-1.645	1.645	0	0	**0**	26	0.000
S1o	-1.645	1.645	12	1	**13**	943	1.379
S2o	-1.645	1.645	0	3	**3**	1145	0.262
B1o	-1.645	1.645	30	29	**59**	3022	1.952
B2o	-1.645	1.645	18	2	**20**	1127	1.775
C1o	-1.645	1.645	62	1	**63**	737	8.548
C2o	-1.645	1.645	0	3	**3**	629	0.477

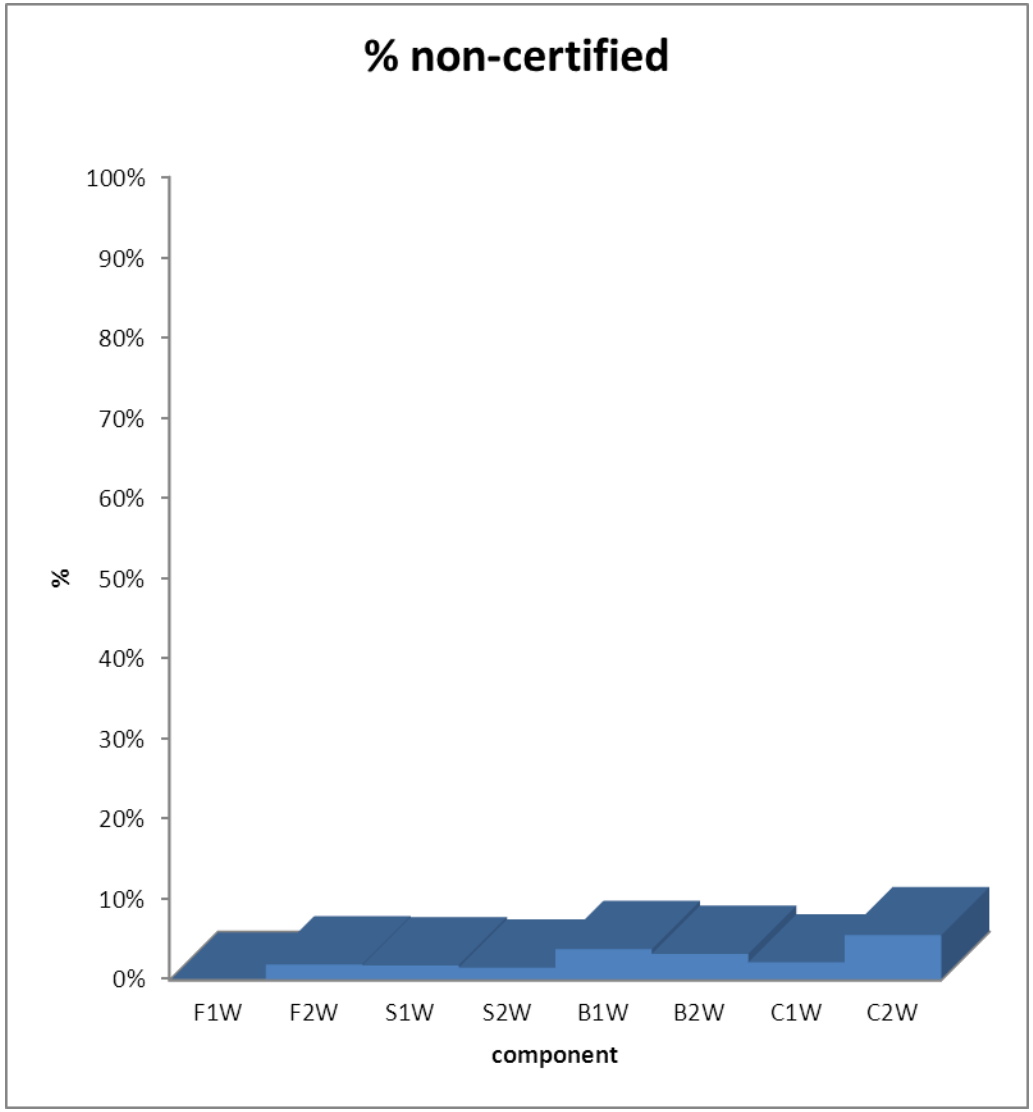
RESULTS 2013: SPEAKING



RESULTS 2013: WRITING

	cut offs low	high	lenient	severe	non-certified count	total analysed	% non-certified
F1w	-1.645	1.645	0	0	**0**	15	0.00
F2w	-1.645	1.645	13	2	**15**	763	1.97
S1w	-1.645	1.645	17	0	**17**	918	1.85
S2w	-1.645	1.645	13	1	**14**	888	1.58
B1w	-1.645	1.645	27	85	**112**	2896	3.87
B2w	-1.645	1.645	0	26	**26**	794	3.27
C1w	-1.645	1.645	6	7	**13**	586	2.22
C2w	-1.645	1.645	0	24	**24**	428	5.61

RESULTS 2013: WRITING



ANALYZING RATERS

1. Live test ratings

- Reliability of the scores assigned by pairs of raters
- Analysis of inter-rater correlation

2. Input

- Excel Spreadsheet with rater, pairings, grade

3. Outputs

- Reliability statistics: Cronbach‘s Alpha by pair
- Correlation: Intraclass Correlation Coefficient by pair
- Measure of Agreement: Kappa, crosstabulation of grades

LIVE-TEST ANALYSIS

Reliability Statistics

Cronbach's Alpha	N of Items
.991	2

Case Processing Summary

		N	%
Cases	Valid	24	10.0
	Excluded[a]	216	90.0
	Total	240	100.0

LIVE-TEST ANALYSIS

Intraclass Correlation Coefficient

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0		
		Lower Bound	Upper Bound	Value	df1	df2
Single Measures	.982	.958	.992	108.395	23	23
Average Measures	.991	.979	.996	108.395	23	23

The ICC assesses rating reliability by comparing the variability of different ratings of the same case (candidate + criterion) to the total variation across all ratings and all cases.

LIVE-TEST ANALYSIS

PAIR4 RATER1 x PAIR4 RATER2 Cross tabulation

Count

		PAIR 4 RATER 2						Total
		3.0	4.0	4.5	6.0	7.5	10.0	
PAIR 4 RATER 1	3.0	**1**	0	0	0	0	0	1
	4.0	0	**3**	0	0	0	0	3
	4.5	0	0	**2**	0	0	0	2
	6.0	0	0	0	**2**	0	0	2
	7.5	0	0	0	0	**1**	0	1
	10.0	0	0	0	0	**1**	**14**	15
Total		1	3	2	2	2	14	**24**

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	**.931**	.068	8.205	.000
N of Valid Cases		24			

NEXT STEPS?

- Exploring the number of necessary samples to improve the validity of the statistical analysis of the rater reliability
- Further development of rating criteria
- Common data collection and analysis procedures for ALTE members

MANY THANKS FOR YOUR ATTENTION.