

*4th Annual **EALTA** Conference
Sitges, 15th - 17th June 2007*

Linking the TestDaF to the CEFR: The Case of Writing Proficiency

Gabriele Kecker & Thomas Eckes, TestDaF-Institut, Germany,
gabriele.kecker@testdaf.de, thomas.eckes@testdaf.de

Objectives

- TestDaF – CEFR:
Validating the claim of link through an empirical approach
- Following the methodological steps as outlined in the Manual (Council of Europe, 2003)
- Linking 4 TestDaF subtests (2 receptive skills and 2 productive skills)
- Here: the case of TestDaF Written Production

TestDaF Levels and the CEFR

Common European Framework of Reference (CEFR)					
A Basic User		B Independent User		C Proficient User	
A1	A2	B1	B2	C1	C2

TestDaF Levels	TDN 3	TDN 4	TDN 5
----------------	--------------	--------------	--------------

TDN – TestDaF-Niveaustufe

TestDaF Written Production



- One task to evaluate the candidate's proficiency at the levels TDN 3, 4, 5
- Examines writing skills that are relevant in an academic context like describing and arguing
- **Objective:** to write a coherent, well-structured text with reference to a topic, giving a clear description of statistical data in a graph or table and developing an argument systematically and comprehensibly

Standardisation of Judgements

Familiarisation
Training



Training in assessing standardised
performances samples



Benchmarking local
performance samples to
CEFR levels



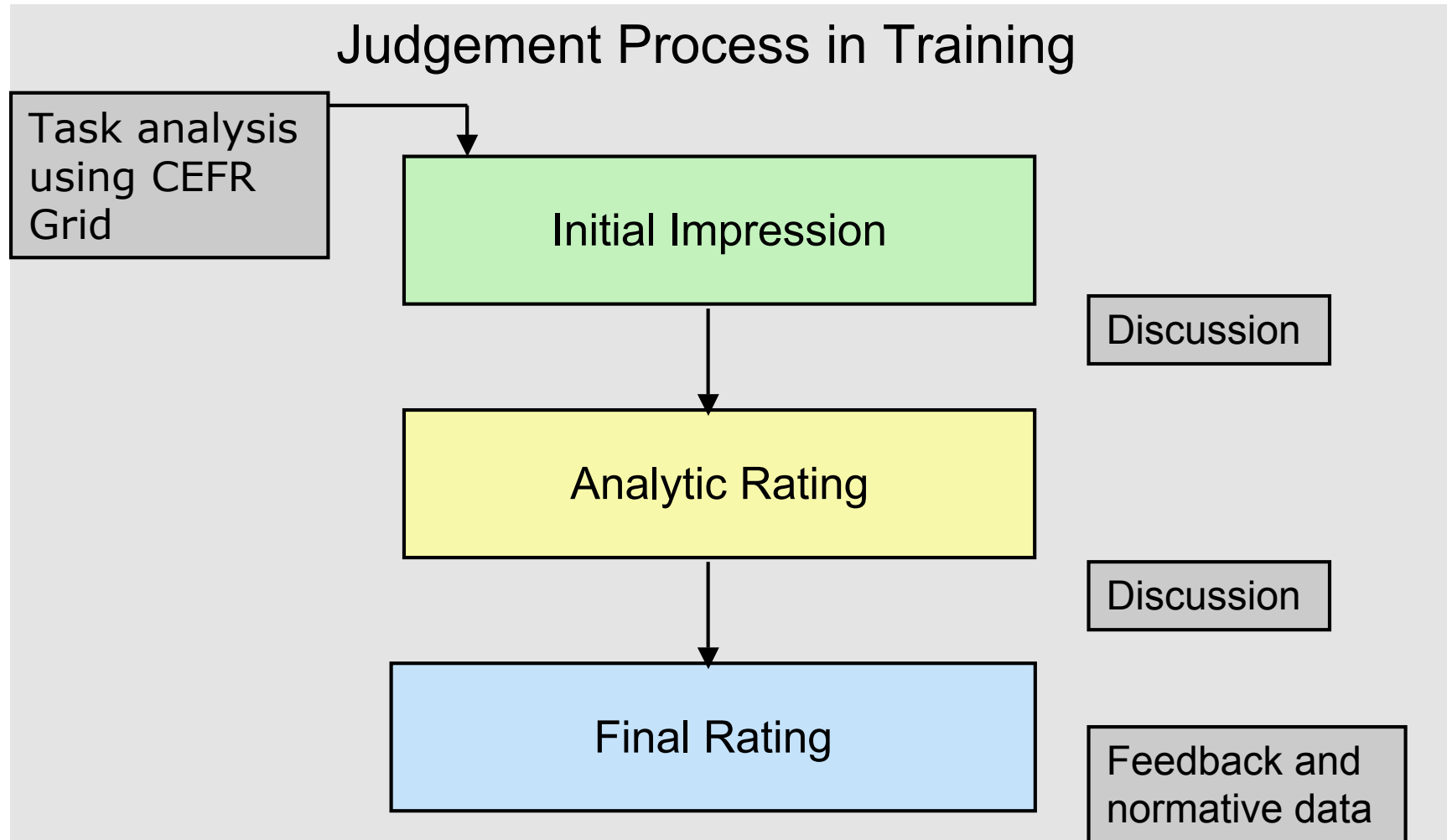
Claim of link to CEFR

Standardisation of Judgements

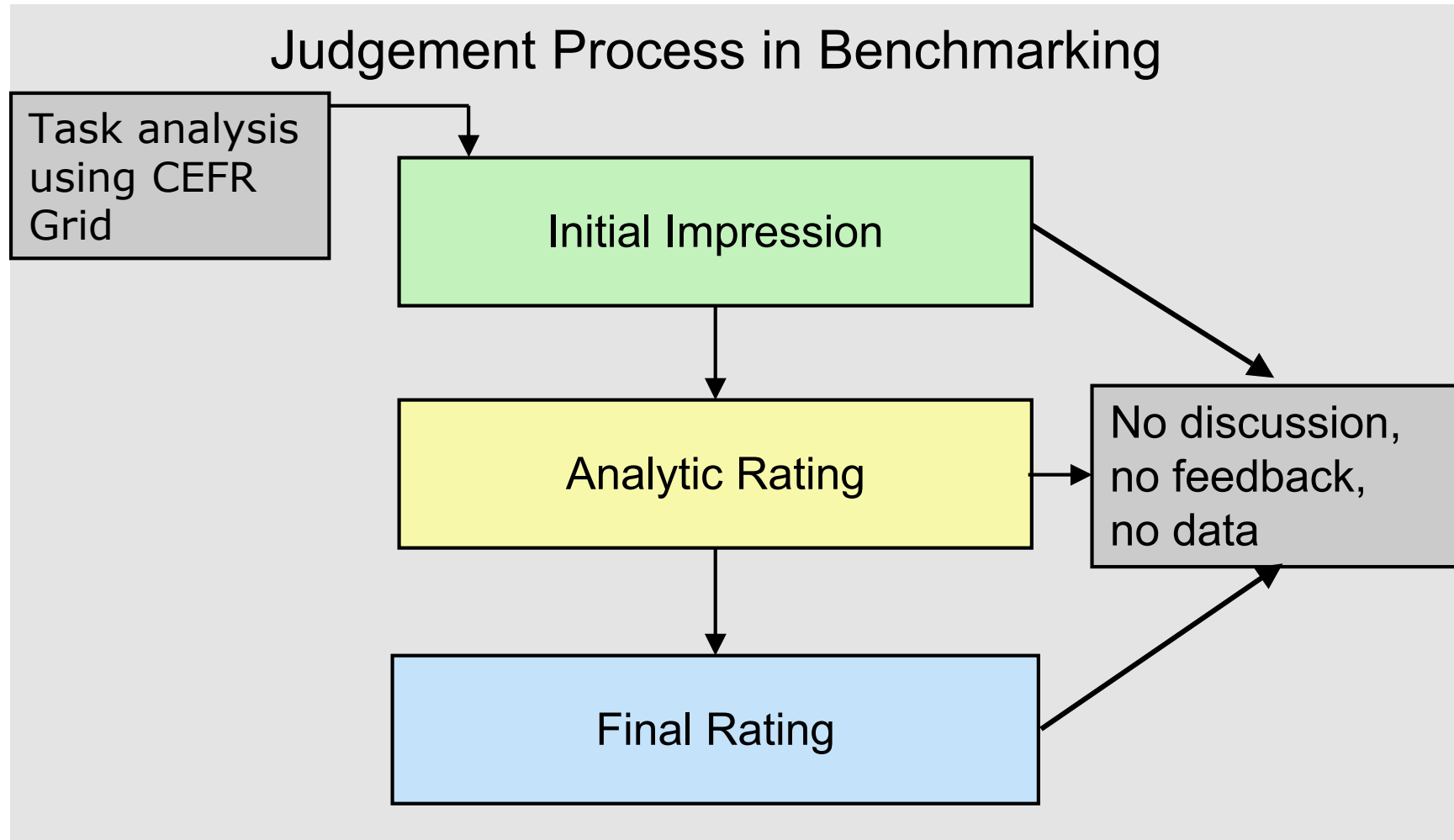
Benchmarking workshop: 2½ days, 14 judges

- Half a day of familiarisation with the CEFR scales; introduction to the CoE project and to the purpose of the benchmarking procedure
- One day of training with standardised samples of the Council of Europe (1 per level available)
- One day of benchmarking TestDaF samples (3 samples per level)

Standardisation of Judgements



Standardisation of Judgements



Standardisation of Judgements

Design Features

- Participants: 14 judges
 - Training stage: 5 standardised samples (CoE)
 - Benchmarking stage: 9 local samples (TestDaF)
- Initial impression, analytic rating, final rating
- Nine-point CEFR rating scale (A1, A2, A2+, etc.)

Standardisation of Judgements

Data Analysis

Interrater reliability and agreement

Many-facet Rasch analysis

Standardisation of Judgements

Interrater Reliability and Agreement

Index	Training
Pearson r (Mean)	.98
Kendall's W	.97
Cronbach's Alpha	.998
Rater Agreement Index (RAI)	.95
Within-Group Agreeem. Index (r_{wg})	.97

Standardisation of Judgements

Summary

- High degree of correctness in final level assignments
- Rater agreement indices consistently high

Conclusion:
Training stage was successful.

Standardisation of Judgements

Interrater Reliability and Agreement

Index	Training	Benchm.
Pearson r (Mean)	.98	.81
Kendall's W	.97	.81
Cronbach's Alpha	.998	.980
Rater Agreement Index (RAI)	.95	.89
Within-Group Agreement Index (r_{wg})	.97	.83

Standardisation of Judgements


Many-Facet Rasch Analysis

- **Final ratings (benchmarking stage)**
 - Most raters acting as independent experts
 - Rater separation reliability = .69
 - Homogeneity statistic: $\chi^2(13) = 46.1$ ($p < .01$)

Benchmarking

TestDaF Samples Final ratings

TDN Level




-  TDN 5
-  TDN 4
-  TDN 3
-  b. TDN 3

Measr	Sample	-Rater	Scale
+ 5 +	<i>high</i>	+ +	(9) +
	Loc_5		
+ 4 +		+ +	8
+ 3 +	Loc_9	+ +	
	Loc_6		
+ 2 +	Loc_2	A5	7
	Loc_8		
+ 1 +		A3	6
		H1	
	Loc_1	H2	
* 0 *	Loc_4	A4 H3	5
		A1	
		H4 T3	
		A2	
+ -1 +		T4	4
	Loc_7	T1	
	Loc_3		
+ -2 +		+ +	3
+ -3 +	<i>low</i>	+ +	(1) +

Benchmarking

TestDaF Samples Final ratings

TDN Level

-  TDN 5
-  TDN 4
-  TDN 3
-  b. TDN 3

Measr	Sample	-Rater	Scale	
+ 5 +	<i>high</i>	<i>severe</i>	(9)	
	Loc_5			
+ 4 +				
			8	C1
+ 3 +	Loc_9			
	Loc_6			
+ 2 +				
	Loc_2	Loc_8	A5	7
				B2+
+ 1 +				
		A3		
		H1		6
				B2
		H2		
* 0 *	Loc_4	A4	H3	
		A1		*
				5
		H4	T3	
		A2		
+ -1 +				
		T4		
		T1		
				4
	Loc_7			
	Loc_3			
+ -2 +				
				3
				A2+
+ -3 +	<i>low</i>	<i>lenient</i>	(1)	

Benchmarking

TestDaF Samples Final ratings

TDN Level



Measr	+Sample	-Rater	Scale	
+ 5 +	<i>high</i>	<i>severe</i>	(9)	
	Loc_5			
+ 4 +				
			8	C1
+ 3 +	Loc_9			
	Loc_6			
+ 2 +	Loc_2	Loc_8	A5	7
				B2+
+ 1 +		A3		
		H1		6
		H2		
		A4	H3	
* 0 *	Loc_4	A1		* 5 *
		H4	T3	
		A2		
+ -1 +		T4		
	Loc_7	T1		
	Loc_3			4
				B1
+ -2 +				
				3
				A2+
+ -3 +	<i>low</i>	<i>lenient</i>	(1)	

↑ Placed too low –
problem with
accuracy criterion?

(Low level of
accuracy in
Samples 2 and 7)

Summary

- **Rater Agreement Statistics**
 - Moderately high agreement in final level assignments
 - Sufficiently high rater agreement indices
- **Facets Analysis**
 - Raters acting as independent experts
 - Sufficient congruence between TDN and B2/C1 levels
 - Most difficult (and problematic) criterion:
Grammatical Accuracy

Conclusion: Claim of link to the CEFR substantiated

Problematic issues

- **Lack of fit between standardised CoE samples and TestDaF samples**
 - Differences in test format
 - Differences in focus on accuracy criterion
- **Lack of fit between TestDaF construct and CEFR descriptors**
 - CEFR lacks a concept for task completion
 - CEFR lacks a descriptor for describing a graph or diagram
- **Different functions of plenary discussion at the Training and Benchmarking stages**

Thank you.

e-mail: gabriele.kecker@testdaf.de
thomas.eckes@testdaf.de
web: www.testdaf.de