

Cambridge, 6-7 December 2007

Reflections on using the Draft Manual: The TestDaF Study

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)

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 linking activities

- **Familiarisation:** In internal meetings team discussions on scale interpretation and exercises
- **Familiarisation:** Exercises for experts/judges during standard-setting and benchmarking workshops
- **Specification:** Work groups of internal staff members working on forms A1-23

TestDaF linking activities

Workshops:

- **Preliminary standard-setting** workshop for receptive skills 06/2005 (10 judges)
- **Benchmarking** workshop for written production in 02/2006 (14 judges)
- **Standard-setting** workshop for receptive skills 05/2006 (15 judges)
- **Benchmarking** workshop for spoken production 10/2006 (12 judges)

TestDaF linking activities

Internal Validation

- Pre-testing of test items (CTT, IRT, many-facet Rasch analysis)
- Test equating using anchor items (common item equating)
- Statistical analysis of rater behaviour
- Validation studies on rater effects, rater types and rater strategies

TestDaF linking activities

External Validation (work in progress)

- Correlation of TestDaF with **teacher judgements** of test-takers' proficiency
- Correlation of TestDaF with **DIALANG**:
Focus on Reading and Listening
Comprehension (Writing constructs
differ strongly)

The TestDaF Study:
Standardisation Phase

Design Features: Receptive Skills

- Participants: 15 judges
- Samples: 9 CoE Item samples for training, 30 TestDaF Reading items + 25 TestDaF Listening items for standard-setting
- Judgement sessions: Training (pre-discussion, post-discussion), Standard Setting (individual decision making)
- Method: modified Angoff

Data and Analysis

- Ratings
 - 9-point rating scale:
A1, A2, A2+, B1, B1+, B2, B2+, C1, C2
- Analysis
 - Interrater reliability and agreement
 - Many-facet Rasch analysis

TestDaF Standardisation: Reading

Interrater Reliability and Agreement

Index	Training		Stand. Setting
	Pre-Disc.	Post-Disc.	
Pearson r (Mean)	.90	.94	.67
Kendall's W	.84	.90	.70
Cronbach's Alpha	.990	.995	.964
Rater Agreement Index (RAI)	.91	.93	.92
Within-Group Agreeem. Index (r_{wg})	.88	.93	.91

Many-Facet Rasch Analysis

- **Ratings (standard setting stage)**
 - 3 raters misfitting
 - Rater separation reliability = .81
 - Homogeneity statistic: $\chi^2(11) = 61.4$ ($p < .01$)

TestDaF Standardisation: Reading

Setting Cut-Scores on the Reading Section

CEFR Level	TDN Level	<i>M</i>	<i>SD</i>	Cut-Score
B1+	-	5.67	3.11	5
B2	3	13.33	3.53	13
B2+	4	21.50	2.58	21
C1	5	29.42	0.67	29

Note. Cut-scores were set according to the procedure outlined in the Manual (CoE, 2003, p. 91); i.e., counting the number of items up to each level, averaging over judges, and rounding averages down. Number of raters = 12.

TestDaF Standardisation: Reading

Setting Cut-Scores on the Reading Section

CEFR Level	TDN Level	<i>M</i>	<i>SD</i>	Cut-Score	Pre-set
B1+	-	5.67	3.11	5	-
B2	3	13.33	3.53	13	15
B2+	4	21.50	2.58	21	21
C1	5	29.42	0.67	29	26

Note. Cut-scores were set according to the procedure outlined in the Manual (CoE, 2003, p. 91); i.e., counting the number of items up to each level, averaging over judges, and rounding averages down. Number of raters = 12.

The Rater Dependence Problem

- IRT assumes local independence
- Training aims at a “reliable consensus”, leading to a strong dependence among raters
- Basic distinction: Raters as “independent experts” vs. raters as “scoring machines”
- FACETS models raters as independent experts
- Rasch-Kappa-Index (Linacre, 2006) to assess the degree of rater dependence

Rasch-Kappa-Index (Linacre, 2006)

- $(\text{obs. agr.}\% - \text{exp. agr.}\%) / (100 - \text{exp. agr.}\%)$
- Values close to 0 indicate rater independence
- Values much greater than 0 indicate strong rater dependence (FACETS analysis inappropriate)
- Negative values indicate unmodeled source of variation in the ratings

TestDaF Standardisation: Receptive Skills

Rasch-Kappa-Index Rater Dependence in Facets Analysis

Analysis	Rasch-Kappa
Reading - Standard Setting	.014
Listening - Standard Setting	-.003

Note. Rasch-Kappa-Index as defined by Linacre (2006). Values greater than 0 indicate more rater agreement than expected on the basis of the Rasch model (raters as rating machines vs. raters as independent experts).

Issues concerning Familiarisation

Manual

Self-assessment of workshop judges (ELP grid)

Qualitative analysis of DIALANG self-assessment scales

Comments

For internal familiarisation very useful

For standardisation workshops confusing to use different versions of scales

Issues concerning Specification

Manual

Forms A1 to A23 to be filled in
(test construct, marking, grading, scoring, test analysis)

Comments

Time consuming, but a good self-evaluation and preparation for standardisation workshops.

Purpose:

To provide evidence of internal consistency and validity –
should be stated more explicitly and prominently.

Issues concerning Standardisation

Manual

Modified **Angoff Method**
and **Benchmarking Method**

Comments

Different function of plenary discussion at the training and benchmarking stages

No differentiation between benchmarking conferences and benchmarking of local tests

Issues concerning Standardisation

Manual

Organisation of standard-setting and benchmarking workshops:

Time management

Comments

Standardisation as homework after training?

Use of the CEFR Grids has to be taken into account

Issues concerning Standardisation

Manual

Number of samples:

for training: 9 to 12

for Benchmarking: 8

(Chapter 5.5.1, Table 5.3)

Standardised CoE items

for training: 15

(Chapter 5.7)

Local items for standard-setting: not indicated

Comments

9 samples for 9 levels

8 samples for 1 level of the local exam?

Range of CoE training items and performance samples available insufficient.

Training samples in English?

How many local items?

Issues concerning Standardisation

Manual

Data analysis:
Interrater reliability
Interrater agreement

Comments

Objective:
Reaching consensus with
experts acting as
independent judges during
decision making

Rasch analysis for data
analysis of discussions
during training
inappropriate

Issues concerning Standardisation

- Is it possible to compare exam results of productive skills to CEFR scale descriptors, when a notion of task completion is lacking in the scales?

Issues concerning Standardisation

- What if specification, standardisation and empirical validation result in different level profiles of a test?
 - What if a test institution has applied different standard-setting methods and received different results?
- ➔ There should be some advice in the Manual how to proceed.

Thank you.

e-mail: gabriele.kecker@testdaf.de

thomas.eckes@testdaf.de

web: www.testdaf.de