European Association for Language Testing and Assessment

Coding open responses – some open questions

Rainer H. Lehmann

Humboldt University, Berlin

Fourth Annual Conference in Sitges , Catalunya, 15 June, 2007

Open questions to be discussed

1. **Quality of coding : an old open question still persisting; options for a paradigmatic shift**
2. **Dimensions of text quality: an old open question requiring new analytic approaches**
3. **Generalizability of ability estimates: an old open question rarely discussed**
4. **Writing skills across languages: sparse evidence for tackling a new open question**
5. **The role of the Common European Frame of Reference: on the relationship of scaling *a priori* and *a posteriori***

Open questions to be discussed (continued)

6. The role of automation: emerging open questions

1. Quality of coding : an old open question still persisting; options for a paradigmatic shift

I. Traditional elements, to be perfected

- Multiple readings
- Analytic subscores
- Use of benchmarks
- Objective micro-criteria
- Automated scoring

II. Pradigmatic shift

- Estimating rater idiosyncrasy
- Adjusting scores for rater idiosyncrasy

Step 1: evaluation of model fit and computation of differences between rater-specific means*

TERM 1: rater

VARIABLES			UNWGHTED FIT		WGHTED FIT	
rater	ESTIMATE	ERROR	MNSQ	T	MNSQ	T
1 1	0.938	0.226	0.91	-0.4	0.93	-0.4
2 2	-1.055	0.233	0.80	-1.0	1.03	0.2
3 3	-0.197	0.215	0.78	-1.1	0.83	-1.2
4 4	0.030	0.220	1.41	1.9	1.38	2.1
5 5	0.284*					

Relevant outcomes:

- raters differ significantly in terms of leniency and, more generally, in terms of the properties of their ‚individual scoring metrics‘.
- deviant raters need to be retrained or excluded from the assessment.

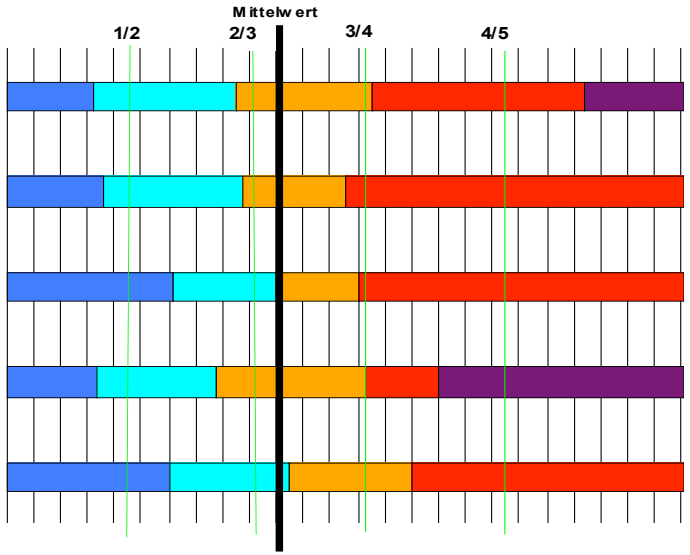
Step 2: computation of position of rater-specific thresholds

TERM 3: rater*step

VARIABLES			UNWGHTED FIT			WGHTED FIT	
rater		step	ESTIMATE	ERROR	MNSQ	T	MNSQ	T
1	1	1	-7.727	0.751	0.74	-1.4	1.50	1.2
1	1	2	-2.443	0.490	0.39	-4.0	0.65	-1.3
1	1	3	2.558	0.475	0.45	-3.5	1.00	0.1
2	2	1	-5.420	0.521	0.52	-2.9	1.06	0.3
2	2	2	-0.306	0.382	1.07	0.4	1.17	1.0
3	3	1	-3.650	0.408	0.67	-1.8	0.85	-0.6
3	3	2	0.066	0.355	0.69	-1.6	1.00	0.0
4	4	1	-6.685	0.582	0.96	-0.1	1.72	1.7
4	4	2	-2.373	0.435	1.77	3.3	1.29	1.2
4	4	3	3.249	0.485	5.15	11.2	2.38	2.6
5	5	1	-4.277	0.449	0.30	-4.8	0.61	-1.9
5	5	2	0.178	0.401	0.56	-2.6	0.93	-0.4

Thresholds for five-point rating scale, five raters

Items (Raters)	N	MEAN
1	52	0,938
2	53	-1,055
3	49	-0,197
4	53	0,03
5	52	0,284



Item Frequencies	N	1	2	3	4	5
rater1	52	7,7	26,9	44,2	19,2	1,9
rater2	53	5,7	32,1	50,9	11,3	0,0
rater3	49	16,3	30,6	34,7	18,4	0,0
rater4	53	7,6	24,5	47,2	13,2	7,6
rater5	52	25,0	36,5	28,9	9,6	0,0

Outcomes

- There are significantly different thresholds for individual raters, requiring adjustments
- It is not yet known (but can and should be investigated) to which degrees such individual tendencies are invariant across tasks and rating dimensions.

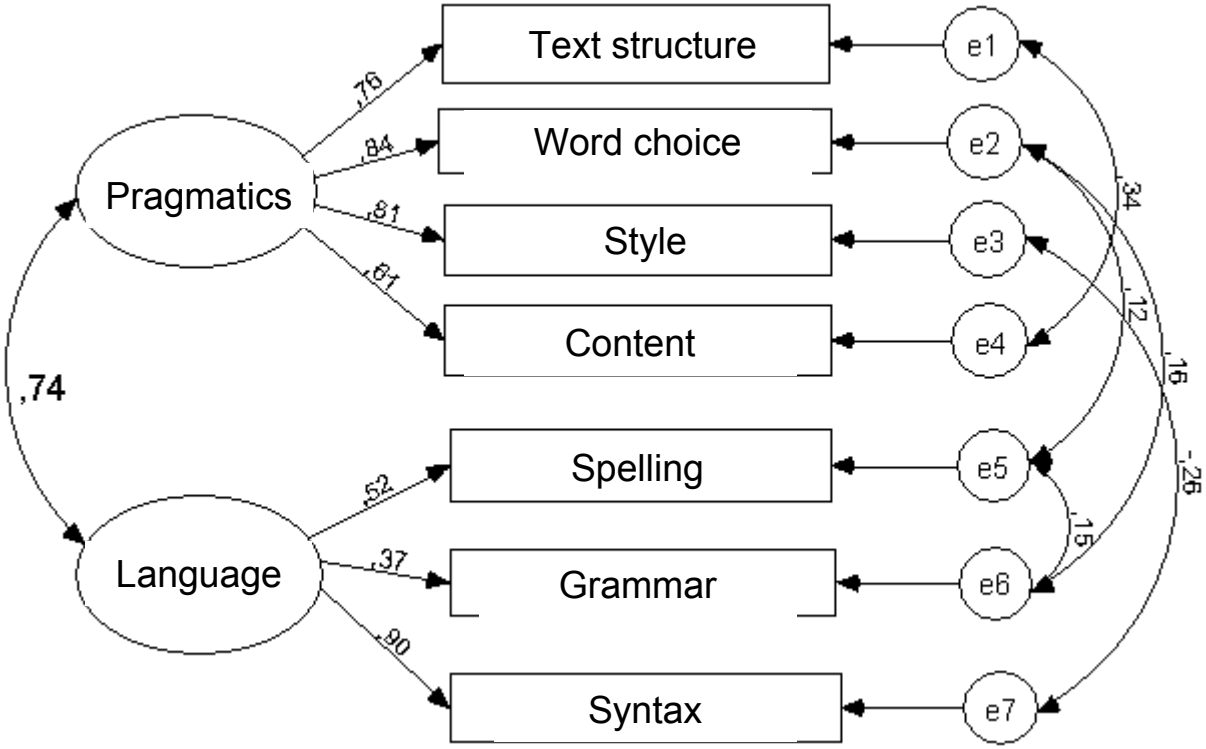
Step 3: Computation of adjusted scores

Rater1	Rater2	score(r1+r2)	score(max)	IRT-Wert	STDDEV
3	4	5,00	9,00	-1,22	1,47
3	4	5,00	9,00	-1,22	1,47
3	4	6,00	9,00	1,44	1,57
3	4	6,00	9,00	1,44	1,57
3	4	3,00	9,00	-5,20	1,64
3	4	3,00	9,00	-5,20	1,64
3	4	5,00	9,00	-1,22	1,47
3	4	3,00	9,00	-5,20	1,64
3	4	8,00	9,00	4,81	1,39
3	4	6,00	9,00	1,44	1,57
3	4	3,00	9,00	-5,20	1,64
3	4	5,00	9,00	-1,22	1,47
3	4	9,00	9,00	6,88	2,02
4	1	6,00	10,00	0,23	1,90
4	1	4,00	10,00	-4,56	1,77
4	1	6,00	10,00	0,23	1,90
4	1	3,00	10,00	-6,71	1,39
4	1	4,00	10,00	-4,56	1,77
4	1	4,00	10,00	-4,56	1,77
4	1	5,00	10,00	-1,93	1,41
4	1	4,00	10,00	-4,56	1,77
4	1	6,00	10,00	0,23	1,90
4	1	5,00	10,00	-1,93	1,41
4	1	6,00	10,00	0,23	1,90
4	1	9,00	10,00	7,06	1,64
4	1	10,00	10,00	5,98	1,49
5	1	4,00	9,00	-2,74	1,57
5	1	6,00	9,00	1,94	1,57
5	1	6,00	9,00	1,94	1,57
5	1				
5	1	4,00	9,00	-2,74	1,57
5	1	2,00	9,00	-7,87	2,19

2. Dimensions of text quality: an old open question requiring new analytic approaches

- Due to high multicollinearity, traditional techniques such as exploratory factor analysis and multi-trait multi-method analysis are likely to fail.
- Structural equation models promise insight into ‚true' dimensional structure.

Structural equation model of dimensionality of text production



3. Generalizability of ability estimates: an old open question rarely discussed

- As a rule, both local evaluations and large-scale assessments operate, for the purpose of obtaining diagnoses in the domain of text production, on very small numbers of tasks per student.
- There are indications of a substantial amount of within-student * between-task variation. Research on this issue appears to be scarce.
- It is highly desirable therefore to obtain sound evidence on the basis of which related questions can be discussed.

4. Writing skills across languages: sparse evidence for tackling a new open question

- International achievement studies have, for understandable reasons, mostly neglected the study of text production, both in the language of instruction and in the foreign languages, let alone investigated simultaneously performance in both domains.
- Respective evidence (from the German DESI study) may be of interest therefore.

Underlying model of language skills

Modality	Auditive		Visual	
Process				
Production	Speaking		Writing	
	Grammar	*Vocabulary*	*Grammar*	*Vocabulary*
	Pronunciation		Spelling	
Reception	Listening		Reading	
	Grammar	*Vocabulary*	*Grammar*	*Vocabulary*
	Speed of perception		Reading speed	

Intercorrelations of language skills German * English

		English (student level)					Class level
		Listening	Reading	Grammar	C-Test	Text production	*Reception*
German (student level)	Vocabulary	0.23	0.29	0.16	0.30	0.24	
	Reading profic.	0.24	0.31	0.18	0.32	0.27	
	Argumentation	0.15	0.24	0.13	0.26	0.28	
	Grammar	0.23	0.29	0.23	0.42	0.36	
	Spelling	0.17	0.26	0.26	0.43	0.34	
	Text production Pragmatics	0.13	0.13	0.13	0.28	0.37	
	Text production language	0.13	0.21	0.18	0.33	0.27	
Class level	*Reception*			0.93	0.95	0.95	0.93
	Production			0.93	0.96	0.96	0.93

Outcomes

- General language tests such as the modified Cloze-Test (C-Test, © Grotjahn) tap language competencies which are relatively stable even across language barriers.
- As opposed to this, measures of writing content or argumentative ability appear to be relatively unstable across languages.
- This underscores the observation of task specificity of mothertongue text production.

5. The role of the Common European Frame of Reference (CEFR): on the relationship of scaling *a priori* and *a posteriori*

- The CEFR presents the challenge of projecting scores and/or ratings onto an *a priori* defined hierarchy of language skills/abilities.
- Expert judgements (e.g., Angoff ratings) are proxies with uncertain theoretical substance.
- Experience so far (based, for instance, on multiple regression estimates) suggests that empirically derived, defensible relationships between the two modes of scaling are not yet above the horizon.

6. The role of automation: emerging open questions

- Despite amazing progress in terms of developing algorithms for automated rating systems, the respective validities require meticulous validation.
- According to the current state of the art, it may be easier to have computers recognize correctly spoken language than to produce electronically dependable and evaluable transcripts of student texts.
- It may also be, however, that PC-based error-tracing routines are already or will soon be producing more valid measures of the writing ‚mechanics' (grammar, spelling).

European Association for Language Testing and Assessment

Thank you very much for your attention and a delightful stay under the Catalan sun!

Fourth Annual Conference in Sitges , Catalunya, 15 June, 2007