Variously rating oral exams: NS, NNS, and more
外/母国語話者などによる口頭試験評価

Rudolf Reinelt

Reference data:
Reinelt, Rudolf (2009) Variously rating oral exams: NS, NNS, and more. In: Reinelt, R. (Ed.) (2009) Into the Next Decade with (2nd) FL Teaching Rudolf Reinelt Research Laboratory EU Matsuyama, Japan, p. 309-336.

contact address
Ehime University Institute for Education and Student Support
Center for General Education
790-8577 Bunkyo-cho 3, Matsuyama-shi
reinelt@iec.ehime-u.ac.jp
愛媛大学　教育・学生支援機構
共通教育センター
790-8577 松山市文京町3　T/F 089-927-9359

Reinelt:Rating Oral Exams: NS, NNS, etc.
外/母国語話者などによる口頭試験評価

Abstract

Aiming at optimizing the practicality of the author's Ehime university German first year courses' oral examinations, this study explores whether raters with different linguistic and other backgrounds, such as native speakers, exchange students, etc. can maintain the required high correlation levels. The results will be relevant both for more easily administering such examinations and for improving the test criteria and the reasoning for the very existence of such courses.

Structure overview

Structure of this presentation

1. Background

2. Rating types and criteria

3. The rater problem: availability

3. 1. Rater types

3. 2. Configuring rater comparability

3.3. Comparing Strictness Values (SV)

3.4. Using SV for correlating

4. Conclusions: comparison results and practical relevance

5. Selective References

Background

1. Background

- of the students in the study

6 y of English in JHS, HS > other FL at university

- of the course:

German > conversation (requested by the students in the first lesson questionnaire)

Class contents available from the author on request, also in Reinelt (2008)Ex post facto Kurrikulum, Ehime U Memoirs of Law & Letters.

- of the exam

final test: Oral exam + writing (administered at the same time in adjacent rooms)

- of the theoretical approach

developing an oral exam for E FL teaching in Japan Jeffrey (n.d.) and Smith & Nederend (1998)

previous literature > papers by the author during development of this German test available from the author on request.

The development of the German elementary course oral exam (1)

1) RR + 1 student

(- ideal: one student to one native speaker

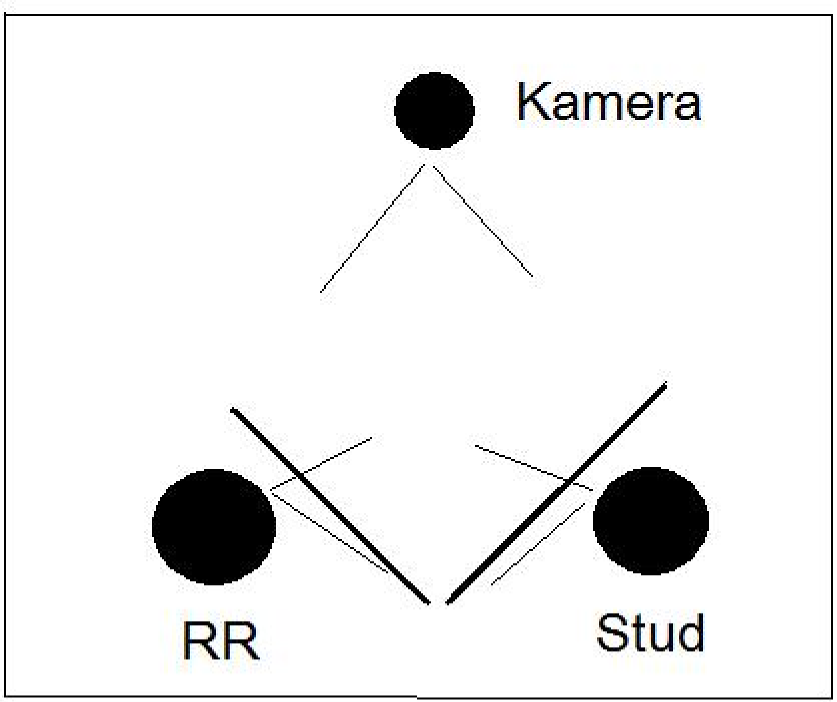
- this impossible due to lack of German NSs)

Rating: only holistic

Various disadvantages

-Criteria (objectivity, validity, reliability ?

- equality, fatigue



RR and one student speak for about 2 mins. Video recording as proof

The development of the German elementary course oral exam (2) (the emergence of raters)

1) RR and 2 students

For reasons of objectvity and practicality: change to the following format:

RR as teacher and rater in one person and two students facing each other

- two (sometimes three) students speak to each other in German for 3 min. as test

- video recording (for later confirmation)

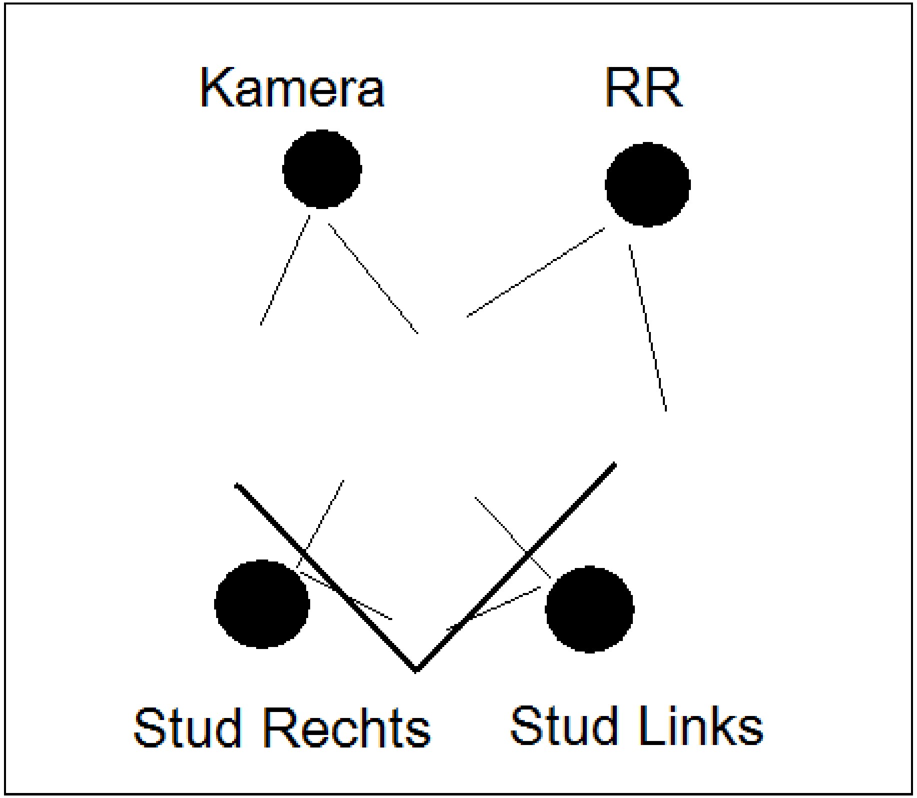
- location (seen from the rater/camera)

This generates the possibility, and need for raters as different rating types become possible

-RR organizes test

-Rechts student on the right

-Stud Links student on the left



2.Rating types and criteria

Two practical types of rating oral exams in university FL courses:

Holistic and criterion-referenced

Holistic

: from experience, but many drawbacks

-simple, grasps overall situation better

-Easily adjustable to the 100 point scale for Japanese university courses

-E.g. Ehime university :

-90 – 100 excellent

-80-89 very good

-65-79 good

-60-64 acceptable

criterion referenced:

limited number of criteria (seven at most, usually 4 to 5 =the highest number one can judge simultaneously (of scorable criteria)

➢scoreboard (see below)

➢Jeffrey & others point out:

➢**Both types are necessary for a good evaluation of an oral exam**

➢However: One rater = one rating type

With RR as only rater > only one rating style possible:

-- either criterion-referenced

-- or holistic

Criterion referenced rating

Need to develop own system for 2FL German

The scoreboard on the right was developed in accordance with Jeffrey

scoreboard for this university (Reinelt 2007)

Weighing of criteria variable, e.g. as in the brackets

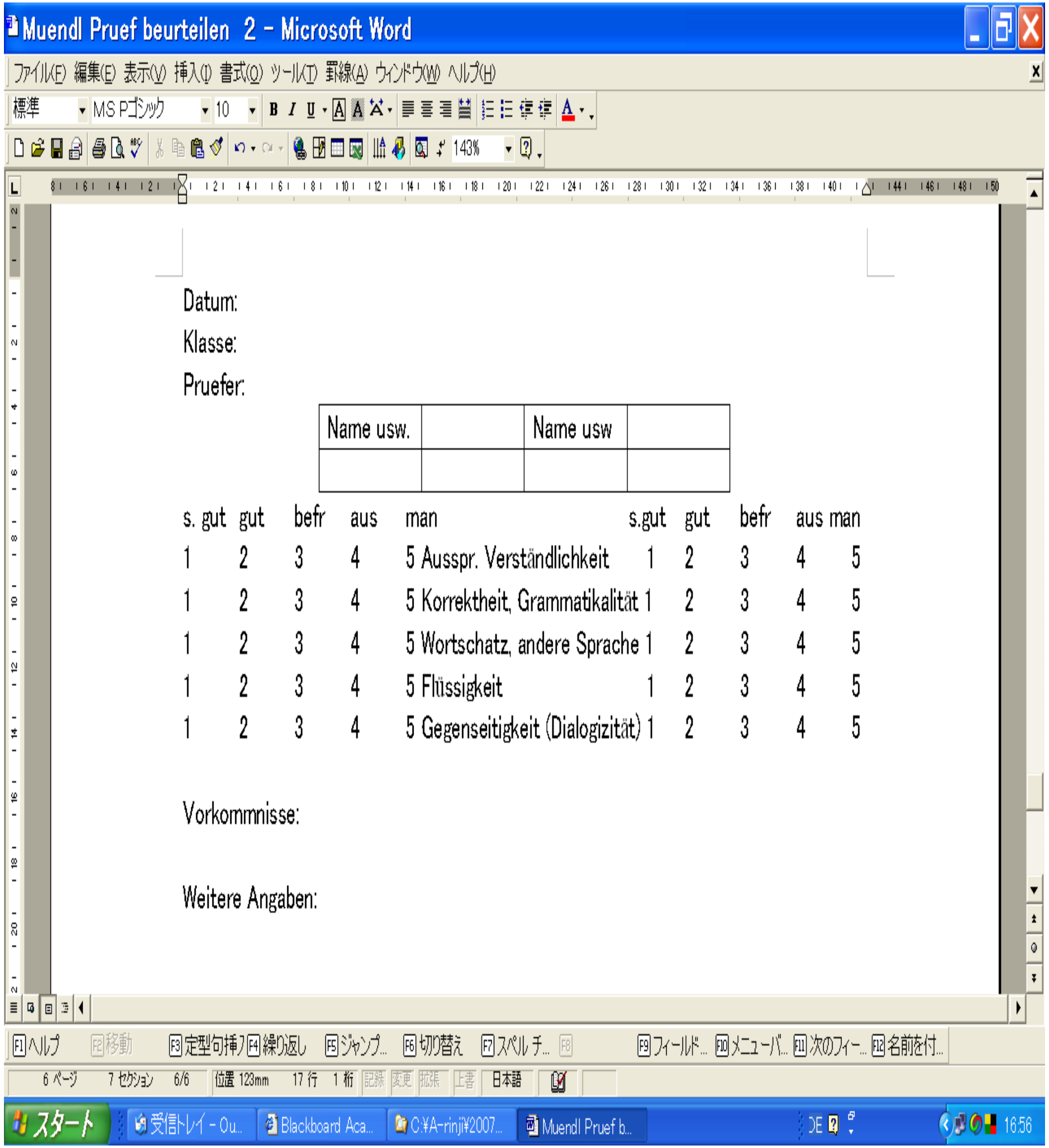
Aussprache = pronunciation (15%) a)

Korrektheit = correctness, grammar (15%) b)

Wortschatz = (richness in) vocabulary (20%) c)

Fluessigkeit = fluency (35%) d)

Gegenseitigkeit = mutuality, dialogicity (15%) e)



Datum:

Klasse:

Pruefer:

Name usw.		Name usw	

s. gut	gut	befr	aus	man		s.gut	gut	befr	aus	man
1	2	3	4	5	Ausspr. Verständlichkeit	1	2	3	4	5
1	2	3	4	5	Korrektheit, Grammatikalität	1	2	3	4	5
1	2	3	4	5	Wortschatz, andere Sprache	1	2	3	4	5
1	2	3	4	5	Flüssigkeit	1	2	3	4	5
1	2	3	4	5	Gegenseitigkeit (Dialogizität)	1	2	3	4	5

Vorkommnisse:

Weitere Angaben:

The development of the German elementary course oral exam (3): Two students, RR + 1 rater

Testees: 2 students sitting facing each other, speak 2-3 min in German

Raters: RR + 1 rater AS (German native speaker)

- RR (organizes the exam and at the same time scores): holistic

- AS: Criterion-referenced scoring

Preparation: Given to the rater

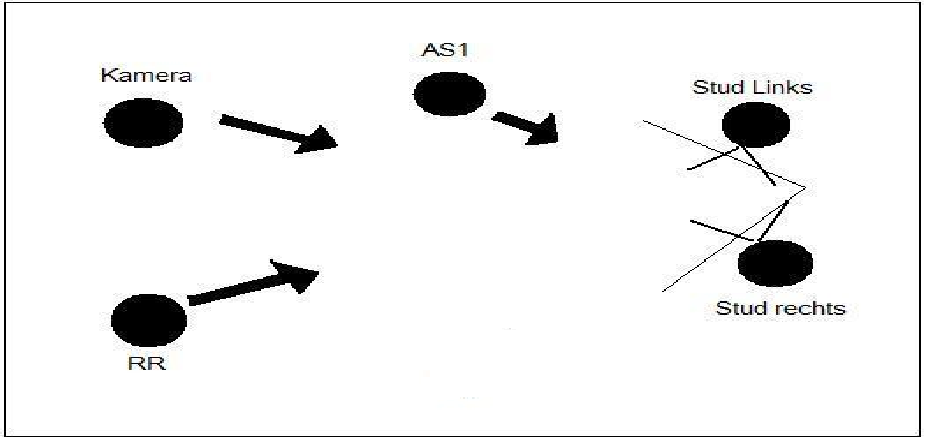
-scoreboard, but no previous training

< own FL (required two to three FL learning in high school in Germany considered as sufficient experience (note: this is only a hypothesis!)

As1 = exchange student

Stud **links** = student on the left

Stud rechts = student on the right



The development of the German elementary course oral exam (4): Two students, RR + 2 raters

Testees: 2 students etc.as above

Raters: RR + 2 raters:

Idea: The more raters the more objectivity, etc.

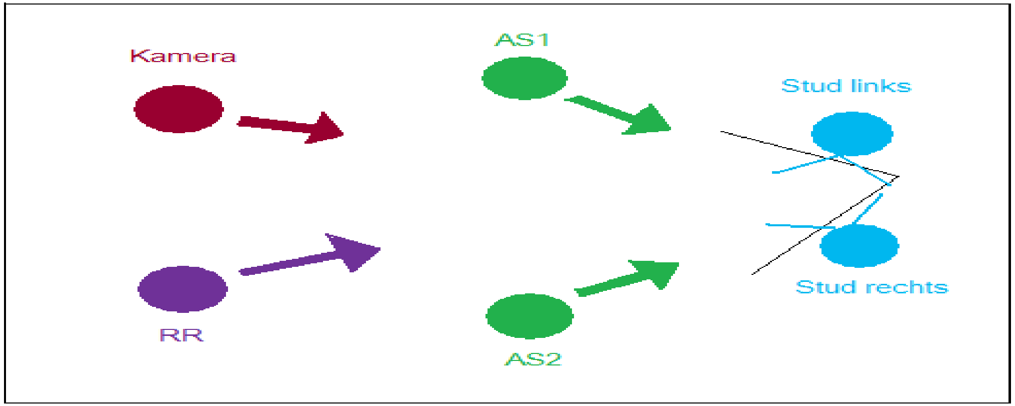
for final results Reinelt (2010)—

RR holistic

AS1 and AS2 separately rate according to the criteria on the scoreboard

As1 = exchange student 1

As2 = exchange student 2



Three raters: Correlation

Three raters:
Example: RR + JP and DV (exchange students)

14 Cases with 3 Raters
Values in the example on the right already weighed and adjusted to the Japanese university system

Interrater correlation machine: Ulm university > use ICC3k

Wanted correlation:
➢90-95% professional
But: For amateur, > 80% satisfactory (Grotjahn 2005)

Correlation in the example on the right:
0.8606538068940285
Not great, but enough for the purposes here

A system with two raters is also possible (but is it really necessary?)

JP	DV	RR
51	50	30
79	77	68
91	100	91
94	100	91
85	70	78
84	73	85
72	90	80
72	97	89
75	89	92
73	73	83
67	64	70
80	71	96
67	63	66

3. The rater problem: availability!

If raters are available, of what kind are they, and how can the correlation be guaranteed?

(raters' majors in brackets)

3.1. Rater types (raters available from Ehime and Matsuyama U German native speaker (NS) exchange students) 2006 through 2009)

- NS professional FL teacher with scoring training RR
- NS professional FL teacher without scoring training KT
- NS exchange students FL related EP (nordic languages), DV (chinese)
- NS exchange students, major not FL related JP (psychology, chemistry), HS (information sciences)
- NNS Chinese exchange student (statistics, with extensive knowledge of German: passed univ entrance exam) ZH
- mother tongue J, + 1 year target language experience in target language country YA

3.2.Rater comparability

3. 2. Configuring rater comparability from interrater correlation (other measures possilbe, but more difficult)

- compare holistic vs holistic (not necessary here and not available in this paper

- compare criterion referenced vs. holistic: see above the three rater example using the crit ref value +%of every criterion relative to the Japanese university scale

- comparing only criterion referenced raters: use: raw data (here only simplest practicable figuring)

For each rater: Sum of all points given (1 to 5) for the 5 criteria for 1 student/5 =average for each student > sum of averages

= average of the sum of averages of all points (1 to 5) given for criteria a to e =

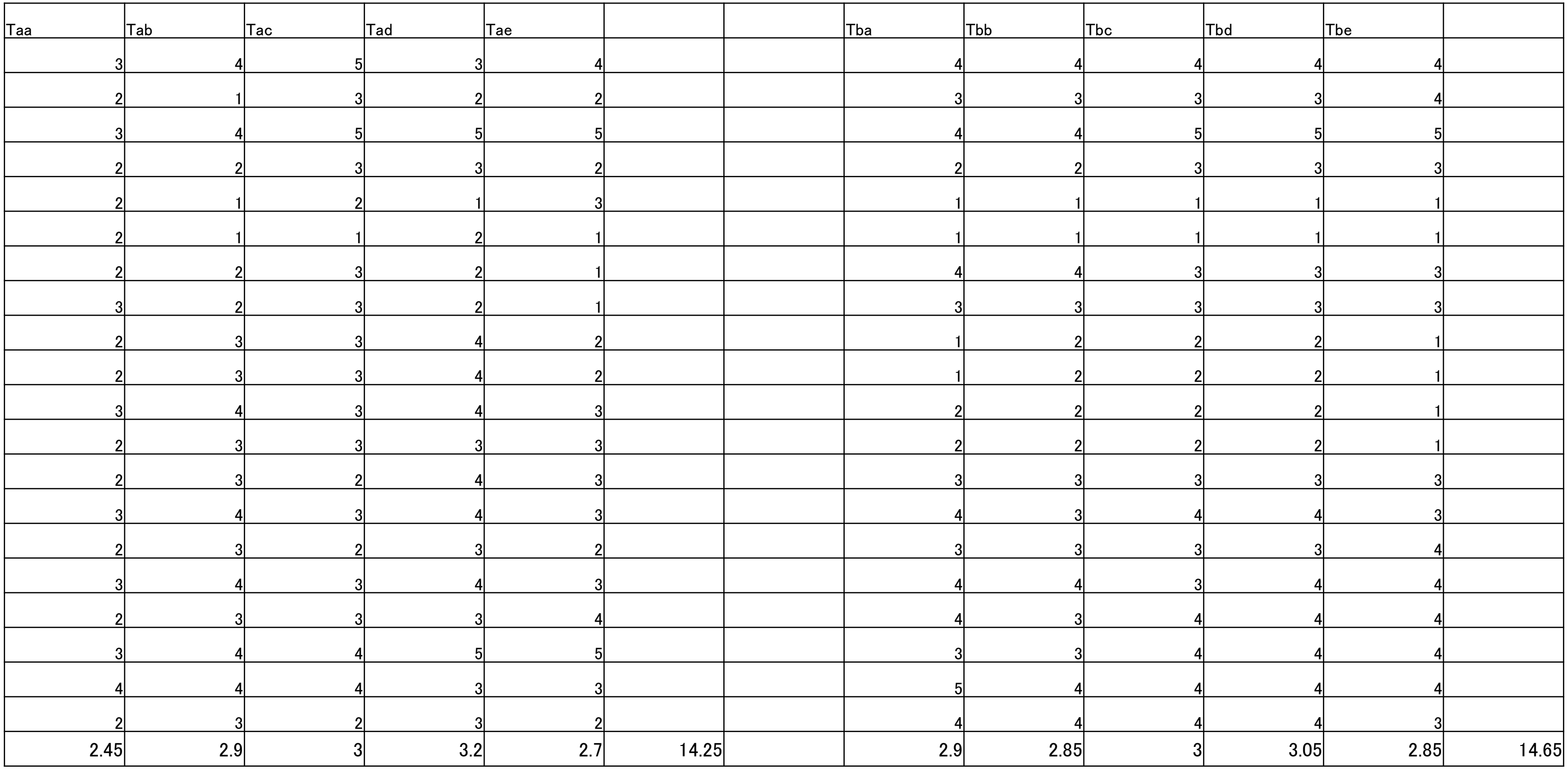
strictness value SV per rater per test

This SV can be used in comparisons of ratings, but it can also be adjusted by later softer or stricter ratings of the rater. Easy configuration.

Example for comparison of any two raters Ta and Tb scoring the five criteria:

3.3. Comparing SVs in one test

3.3. Comparing SVs between pairs of criterion referenced raters in one test



Taa	Tab	Tac	Tad	Tae			Tba	Tbb	Tbc	Tbd	Tbe	
3	4	5	3	4			4	4	4	4	4	
2	1	3	2	2			3	3	3	3	4	
3	4	5	5	5			4	4	5	5	5	
2	2	3	3	2			2	2	3	3	3	
2	1	2	1	3			1	1	1	1	1	
2	1	1	2	1			1	1	1	1	1	
2	2	3	2	1			4	4	3	3	3	
3	2	3	2	1			3	3	3	3	3	
2	3	3	4	2			1	2	2	2	1	
2	3	3	4	2			1	2	2	2	1	
3	4	3	4	3			2	2	2	2	1	
2	3	3	3	3			2	2	2	2	1	
2	3	2	4	3			3	3	3	3	3	
3	4	3	4	3			4	3	4	4	3	
2	3	2	3	2			3	3	3	3	4	
3	4	3	4	3			4	4	3	4	4	
2	3	3	3	4			4	3	4	4	4	
3	4	4	5	5			3	3	4	4	4	
4	4	4	3	3			5	4	4	4	4	
2	3	2	3	2			4	4	4	4	3	
2.45	2.9	3	3.2	2.7	14.25		2.9	2.85	3	3.05	2.85	14.65

Comparing raters

Comparing ***in pairs***:

For five tests we had two raters available: Pairs A B; E F; H J Ta Tb; V X. These were the scores (rat1 and rat2) in the tests

Paare	rat1	rat2
A vs B	10.88462	11.38462
E vs F	11.2	10
H vs J	10.20833	11.625
Ta vs Tb	14.25	14.65
V vs X	9.333333	7.714286
	stdev	2.095524

Comparing ***individual raters***

absolute ranking of raters according to points SV, i.e.

sort SV according to points

Tb	14.65
Ta	14.25
J	11.625
B	11.38462
E	11.2
A	10.88462
H	10.20833
F	10
V	9.333333
X	7.714286

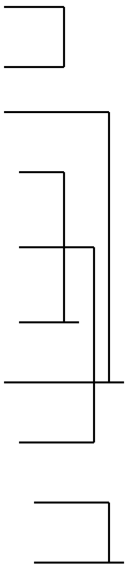
Notes

1) Averages of sums are of a wide range from 7,. to 14

2) Overall differences wide enough to exclude accidental proximity or smilarity

Pair-wise comparison across the five tests

Tb	14.65
Ta	14.25
J	11.625
B	11.38462
E	11.2
A	10.88462
H	10.20833
F	10
V	9.333333
X	7.714286



3.4. Using SV for correlating

3.5. Using SV for correlating

- all raters of one test: in rat1 vs rat2 = pairwise

(here only sums, individuals also possible)

- Standarddeviation to delimit reasonable variation (all within this, less than half)

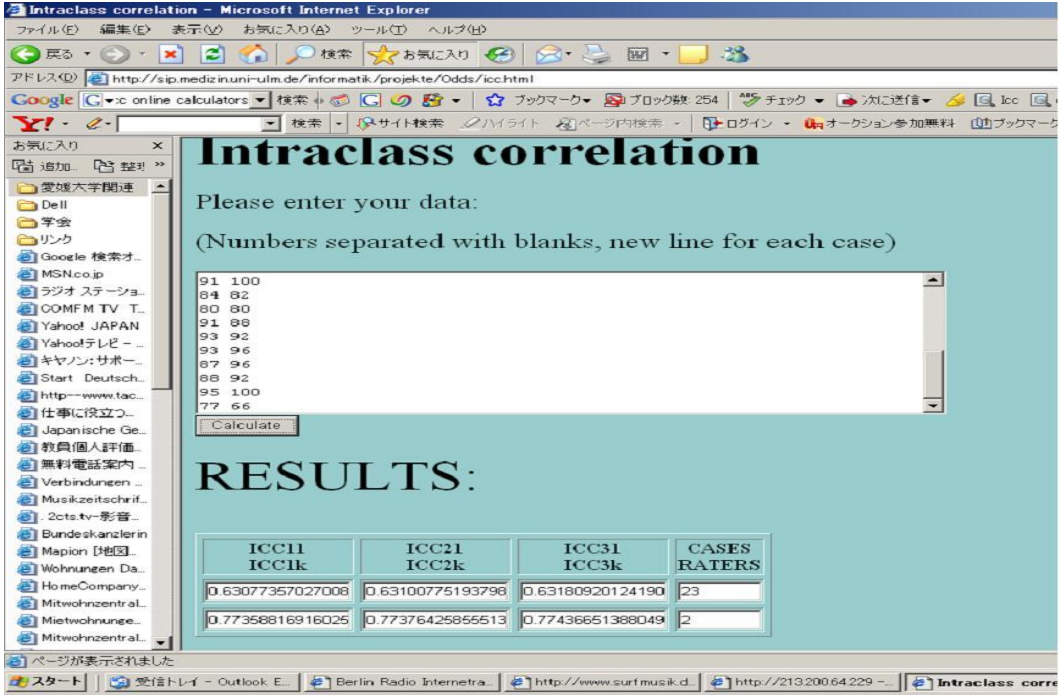
-enter the 5 pairs in Ulm interrater correlation calculator: <http://sip.medizin.uni-ulm.de/informatik/projekte/Odds/icc>

-usually 80% o.k. for non-professional raters, 90% for professionals!

Paare	rat1	rat2
A vs B	10.88462	11.38462
E vs F	11.2	10
H vs J	10.20833	11.625
Ta vs Tb	14.25	14.65
V vs X	9.333333	7.714286
	Standardabweichung	2.095524

Intraclass etc. correlation calculator

<http://sip.medizin.uni-ulm.de/informatik/projekte/Odds/icc>


Intraclass correlation - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
http://sip.medizin.uni-ulm.de/informatik/projekte/Odds/icc.html
お気に入り

Intraclass correlation

Please enter your data:

(Numbers separated with blanks, new line for each case)

```
91  100
84  82
80  80
91  88
93  92
93  96
87  96
88  92
95  100
77  66
```

Calculate

RESULTS:

ICC11 ICC1k	ICC21 ICC2k	ICC31 ICC3k	CASES RATERS
0.63077357027008	0.63100775193798	0.63180920124190	23
0.77358816916025	0.77376425855513	0.77436651388049	2

Interrater correlation between rat1 and rat2

result for the interrater correlation of the five pairs: 0,91 of 1 !

4. Conclusion: Comparison results and practical relevance

If we consider classes = groups of raters

-Both raters' results are fairly similar per class, i.e. the raters a and b are very near each other in their judgments

-(rather than mixed/overlapping except for the center, where all results are comparably close);

-and go in the same direction (thus 14 and 14, and not 13 and 7) vs the extremes, i. e..

- either both go up (bad class in German rating) or both go down (good class);

- overlap only in the middle and less than the standard deviation of all doubly rated tests;

- the interrater correlation of 0.91 almost equals professionals;

- if either of A or B correlates with RR, the other does too sufficiently by equation;

- either rater is reliable/usable;

- although the values are not numerical but nominal, they are still interpretable

- this means that raters usually have a good feeling for a good and for a bad class

- this means that even without training we obtain a good correlation

-this means that no training is necessary

Decoding the raters

Decoding the raters above, we find there are:

No systematic differences discernible!

In any class with an average better than average or average worse than average scoring, and for all scores in the middle, the following holds for the exchange students (and all others) who cooperated in this study: Raters scored pair-wise in the same direction leading, to the following constellations:

Natural science majors are close to literature majors, and long target language experience holders are close to information majors, and raters are also constant along the classes they rated. All depended on the overall tendency of the class result.

DV	Tb	14.65
JP	Ta	14.25
YG	J	11.625
JP	B	11.38462
US	E	11.2
US	A	10.88462
US	H	10.20833
EP	F	10
HS	V	9.333333
An	X	7.714286

Selective References

Intraclass correlation (n.d.) http://sip.medizin.uni-ulm.de/informatik/projekte/Odds/icc.html

Jeffrey, D. (o. J.): The Challenges of Creating a Valid and Reliable Speaking Test as Part of a Communicative English Program http://www.nuis.ac.jp/~hadley/publication/jeffrey/jeffrey-speakingtest.htm

Reinelt, R. (2008)「一年生のドイツ語口頭試験における自然さ」 ”Real Communication aspects in first year German as 2FL oral examination”In: Reinelt, R. (ed.) (2008) Research in Communication and Medicine (Proceedings 第10回日本コミュニケーション学会（CAJ）中国四国支部大会 共催　医療コミュニケーション教育研究セミナー（第２回）, 広島大学, 2007年12月15日), S.23-35.

Reinelt, R. (2008) "ドイツ語口答試験のドイツ語母語話者による評価 Muttersprachlerbeurteilung von Sprechprüfungen im Deutschunterricht", 2007年度日本独文学会中国四国支部学会, Nov. 10. 2007. 中国四国ドイツ文学論集第４１号, 2008年10月31日, S.73-83.

Reinelt, R. (2008) “Inter-rater reliability in native speaker German beginners course‘ oral examinations”, The 33rd JALT IntReinelt, R.(2008b): Inter-rater reliability in native speaker German beginners course’ oral examinations. In Kim Bradford-Watts (Ed.), *JALT2007 Conference Proceedings*. Tokyo: JALT 2007 Proceedings p.1154-1166.

Reinelt, R. (2008)138. “Ex-post-facto Kurrikulum”(事後(後から)のカリキャラム) 愛媛大学法文学部論集 人文学科編 第25号, 2008年9月. S.111-124.

Reinelt, R. (2009) 「未習外国語学習口頭のためのOutsourcing」第56回中国・四国地区大学教育研究会, 鳥取大学, June. 1. 2008. 「第56回中国・四国地区大学教育研究会報告書」, 鳥取大学, May. 31. 2008. S 182.Ausführlich in CAJ ChugokuShikoku Chapter Newsletter 27, S. 7-10.

Smith & Nederend (1998) Smith, A.F.V. & Nederend, W. (1998). Using Oral Interviews at a Junior College. *The Language Teacher.* (http://www.jalt- publications.org/tlt/files/98/apr/smith.html, besucht Feb. 2008).

Thank you very much for your attention

Papers in this context on request from the presenter

Appendix (1): Data resource for the five pair-wise scored tests

A=E=H	Pairs				Averages		
B=Ta	A vs B	10.88462	11.38462		11.13462		
F	E vs F	11.2	10		10.6		
J	H vs J	10.20833	11.625		10.91667		
Tb	Ta vs Tb	14.25	14.65		14.45		
V	V vs X	9.333333	7.714286		8.52381	11.12502	
X		stdev	2.095524			55.62509	

Appendix (2)

B	2.038462	2.5	2.615385	2.5	1.730769	11.38462
E	1.9	2.3	2.2	2.7	2.1	11.2
F	1.5	2.5	1.5	2.7	1.8	10
H	2.125	2.333333	1.833333	2.166667	1.75	10.20833
J	2.25	2.708333	2.5	2.25	1.916667	11.625
Ta	2.45	2.9	3	3.2	2.7	14.25
Tb	2.9	2.85	3	3.05	2.85	14.65
V	2.404762	1.833333	1.880952	1.833333	1.380952	9.333333
X	1.333333	1.809524	1.952381	1.571429	1.047619	7.714286
						11.12502

Future applications

1. In many ways, this study has shown that oral exams of second foreign languages can reliably be put into numbers, even in Japan. High inter-rater correlations hint at high objectivity. This can be a starting point for convincing administrative bodies, that foreign language teaching beyond English can also be evaluated objectively. As such it can be continued or even expanded according to student's wishes or university policies.
2. The native-speaker rater availability has been explored even further by the author. Even without professional equipment, Skype can enable evaluation all over the world (Reinelt study Dec. 2009), and of course in the target language country or environment. Thus other foreign languages are not even at a long distance disadvantage anymore. However, the problem to find persons volunteering as raters remains as a future task.

Thank you very much for your cooperation. Your comments are welcome at

reinelt@iec.ehime-u.ac.jp