Raili Hildén & Marja K. Martikainen

University of Helsinki

Rater feedback on speaking assessment

EALTA Conference 4. – 7. June 2009

Turku, Finland

HY-TALK

Major aims of the project

http://blogs.helsinki.fi/hy-talk/

- To promote teaching, learning and assessment of communicative oral proficiency in foreign languages in general education and at tertiary level by enhancing the quality of the measurement instrument
- To respond to current need due to the introduction of a specific course of oral proficiency in the upper secondary school

HY-TALK

The goal of the project

- To validate the illustrative scales of speaking included in the national core curricula for general education and upper secondary level by trialing a prototype test of speaking.

The conceptual framework

- Validity argumentation scheme for interpretation of the HY-Talk project data (adapted from Fulcher & Davidson, 2007, 164 – 174; Bachman, 2005)
- The use argument is not considered so far, because test performance bears no consequences for the student.

Claim:

- The illustrative scales of descriptors of oral proficiency included in the national core curricula for language education and the tasks designed to measure students´ oral proficiency in general school education in Finland enable sufficiently valid conclusions about their speaking ability.

The purpose of the HY-Talk study

- The validity claim is supported and challenged by warrants and rebuttals regarding
- **relevance**
- **utility**
- (Intended consequences)
- **sufficiency**

Research Question 1 →Utility

1. What is the degree of consistency between raters of a single jury (inter-rater reliability)?

1a. In terms of dimensions of speaking proficiency (overall task performance, fluency, pronunciation, range and accuracy)?

1b. Are there significant differences between raters?

Research Question 2 →Utility

2a. **What is the relation between a rater´s level ratings and verbal comments? (intra-rater reliability)**

2b. **What is the relation between jury ratings and content of subsequent discussion (inter-rater reliability)**

Research Question 3 → Relevance, Utility, Sufficiency

3. What themes do raters introduce when motivating their level ratings?

Research Question 4 →Sufficiency (impact for future development)

4. What features and patterns of interaction emerge in rater discussions?

Context, data and method

- A multimethod approach is adopted to investigate data from multiple sources and of multiple types.

Level scale of the Finnish language curricula

- a proficiency scale was made a part of the new curricula, adapted from the CEFR
- Target levels are specified for the end of grade 6, the end of grade 9 and the end of senior secondary school.
- Scale construction has been investigated by Hildén & Takala (2003)
- Calibration to the CEFR made by Hildén & Takala (2006)
- Texts, themes and tasks selected from the CEFR

Proficiency Scales for language core curricula for general education and upper secondary level (LOPS 2003; POPS 2004)

Taitotaso A1	Suppea viestintä kaikkein tutuimmissa tilanteissa
A1.1	Kielitaidon alkeiden hallinta
A1.2	Kehittyvä alkeiskielitaito
A1.3	Toimiva alkeiskielitaito
Taitotaso A2	Välittömän sosiaalisen kanssakäymisen perustarpeet ja lyhyt kerronta
A2.1	Peruskielitaidon alkuvaihe
A2.2	Kehittyvä peruskielitaito
Taitotaso B1	Selviytyminen arkielämässä
B1.1	Toimiva peruskielitaito
B1.2	Sujuva peruskielitaito
Taitotaso B2	Selviytyminen säännöllisessä kanssakäymisessä syntyperäisten kanssa
B2.1	Itsenäisen kielitaidon perustaso
B2.2	Toimiva itsenäinen kielitaito
Taitotaso C1-C2	Selviytyminen monissa vaativissa kielenkäyttötilanteissa
C1.1	Taitavan kielitaidon perustaso

Tasks

- 3 sets of tasks:
- at the end of year 6
- at the end of year 9
- At the end of upper secondary level

Tasks included in this study:

- Monologic presentation
- At the airport
- At home
- Planning an outing

All prompts and instructions were given in L1 (Finnish)

Data and method of analysis

- A set of speaking tasks were constructed and targeted to the level defined for each check point
- 42 video recorded performance samples from students of German, 32 from students of English
- 5-7 raters in German
- 5 raters in English
- Level ratings (1- 10) given to the samples on four dimensions of speaking proficiency
- Video records of rating sessions (giving reasons for the assignments, but no changes made to single ratings after the discussions)

RQs, data and method of analysis

1. What is the degree of consistency between raters of a single jury (inter-rater reliability)?	Level ratings	ANOVA
2. What is the relation between numeric indicators of a rater´s level assignments and his/her verbal comments? (intra-rater reliability)	Level ratings Video records of rating sessions	ANOVA Qualitative Content analysis
3. What themes do raters introduce when motivating their assignments?	Video records of rating sessions	Qualitative Content analysis
4. What features and patterns of interaction emerge in rater discussions?	Video records of rating sessions	Discouse analysis, interaction analysis

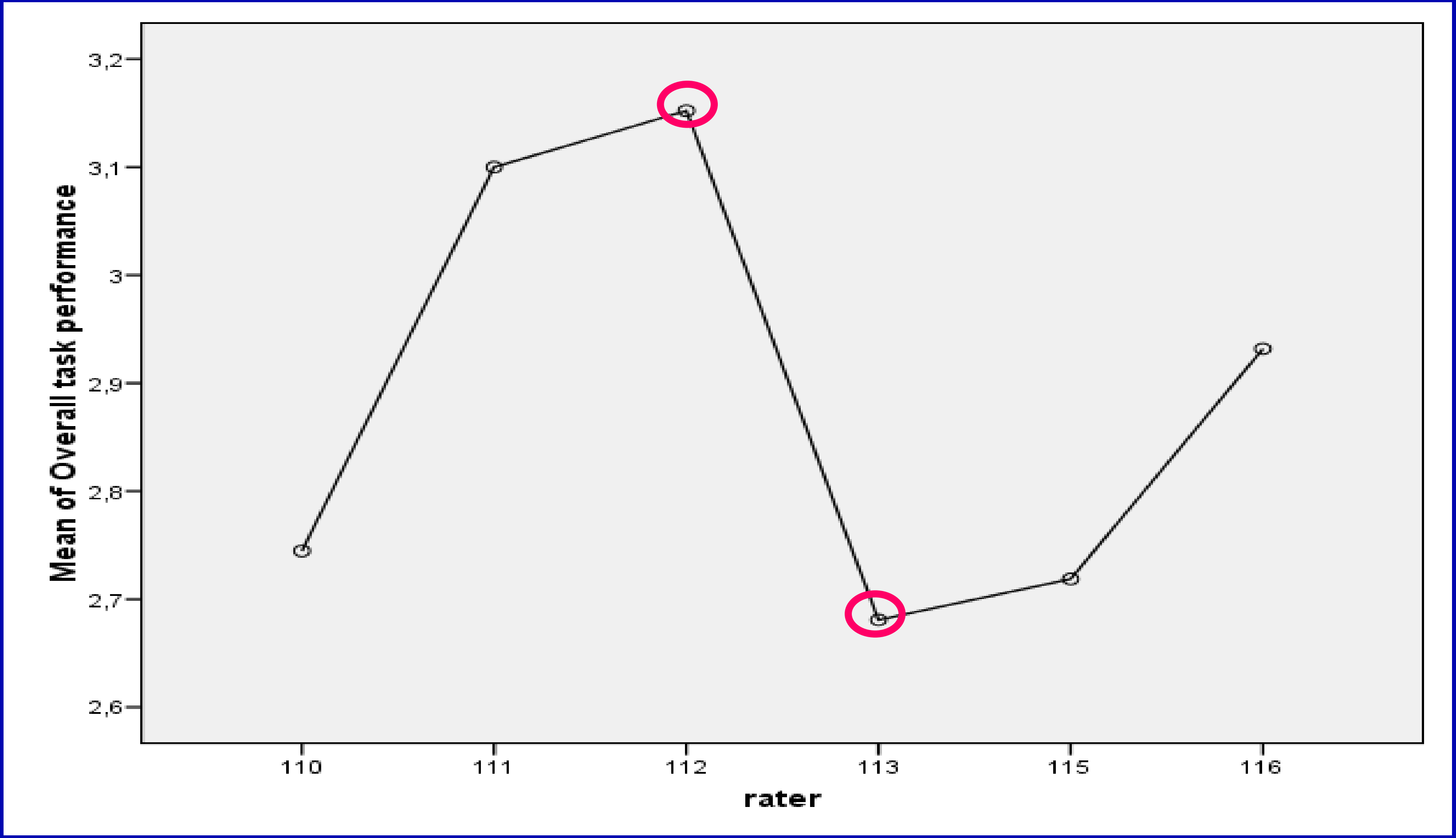
Research Question 1

1. What is the degree of consistency between raters of a single jury (inter-rater reliability)?

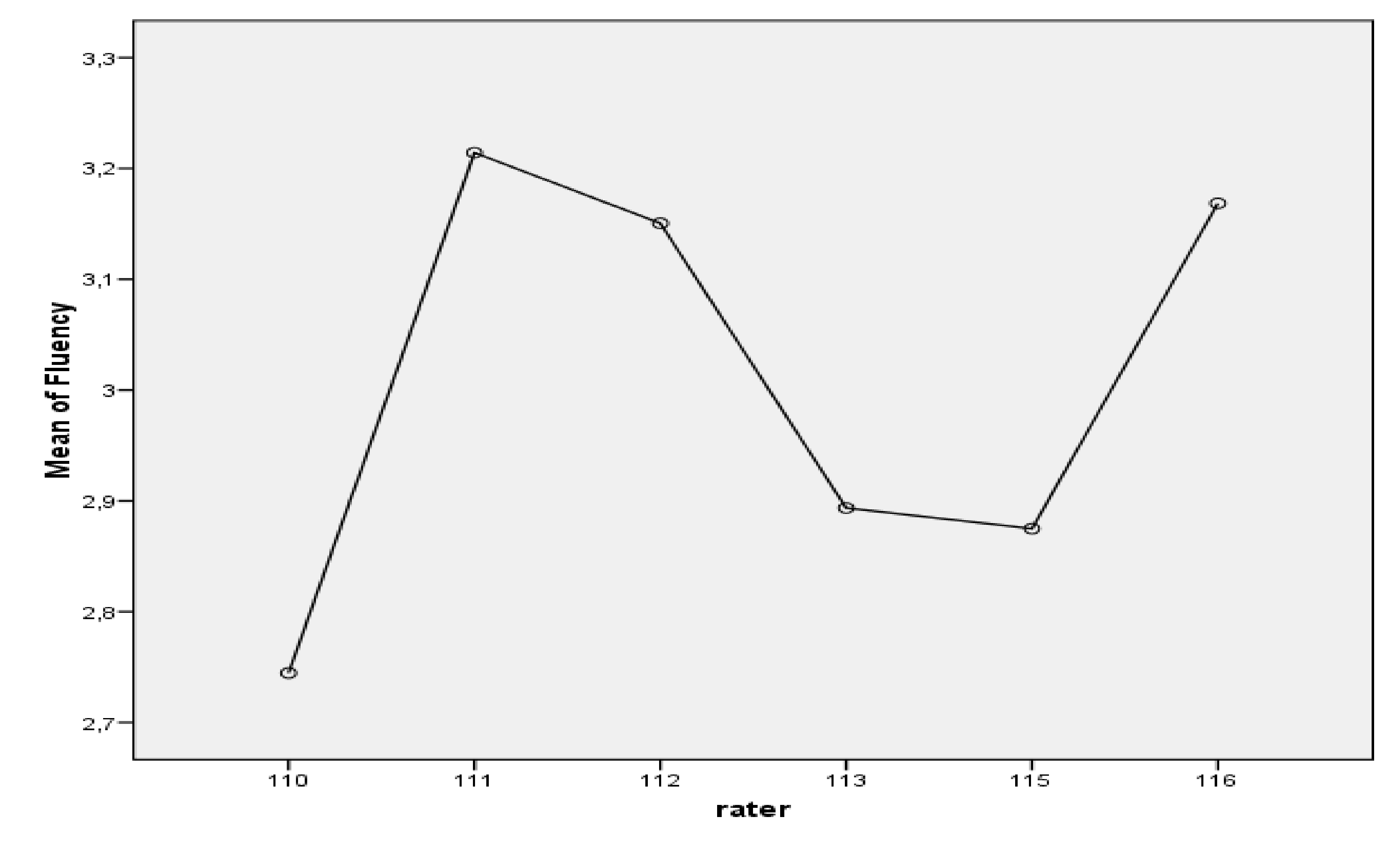
1a. In terms of dimensions of speaking proficiency (overall task performance, fluency, pronunciation, range and accuracy)?

1b. Are there significant differences between raters?

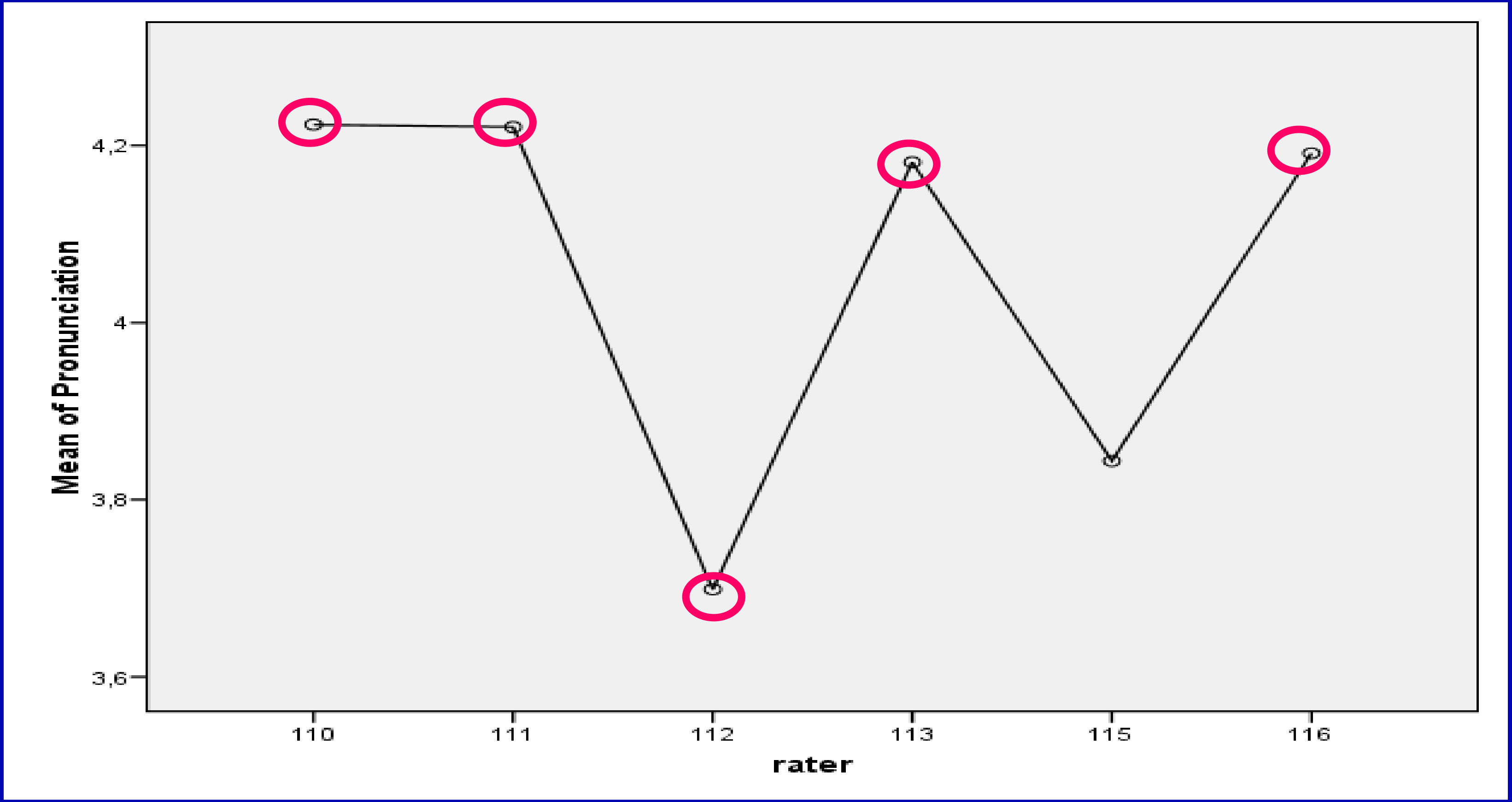
RQ1a: The degree of consistency between raters of the German jury (inter-rater reliability): Overall task performance



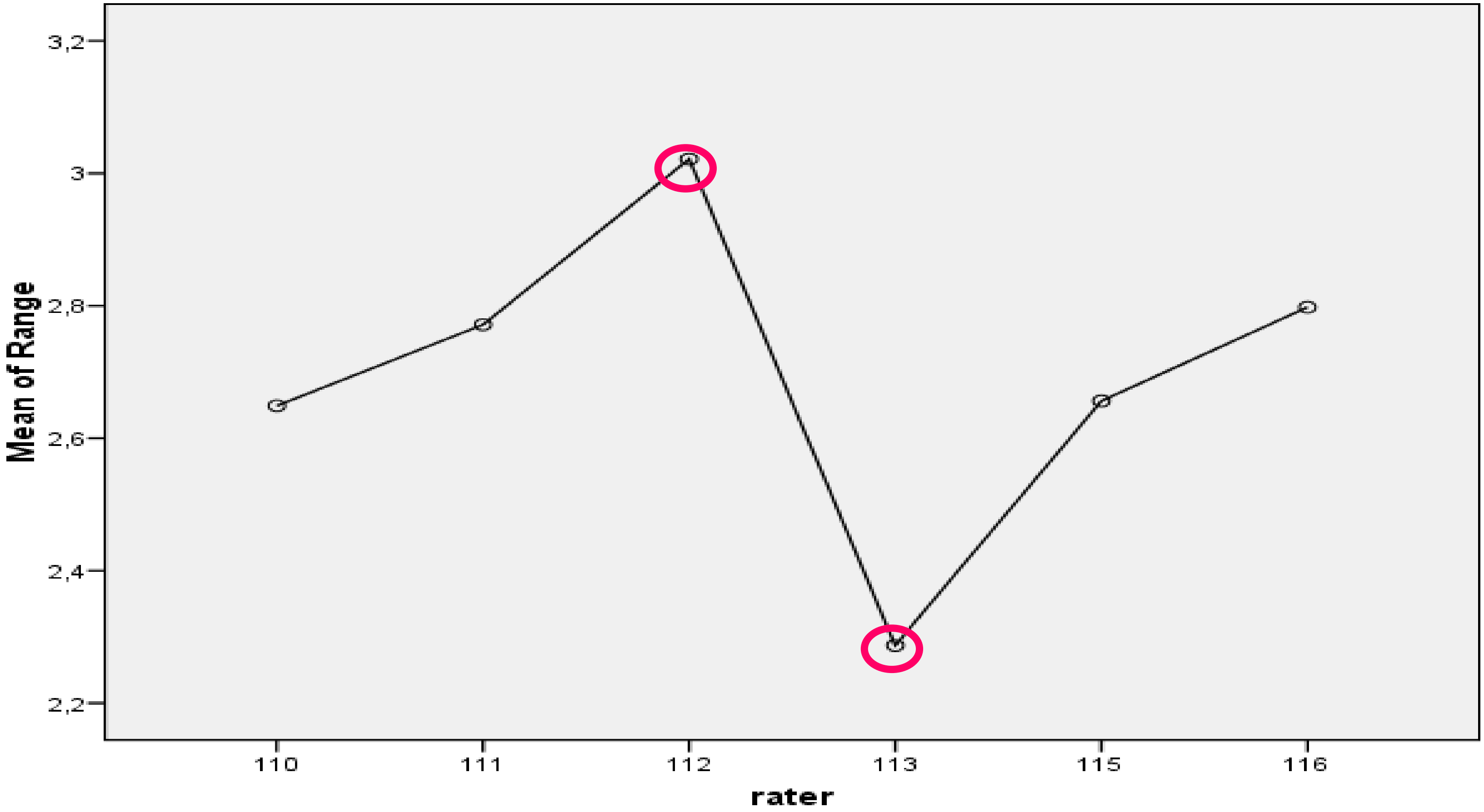
RQ1: The degree of consistency between raters of the German jury (inter-rater reliability): Fluency



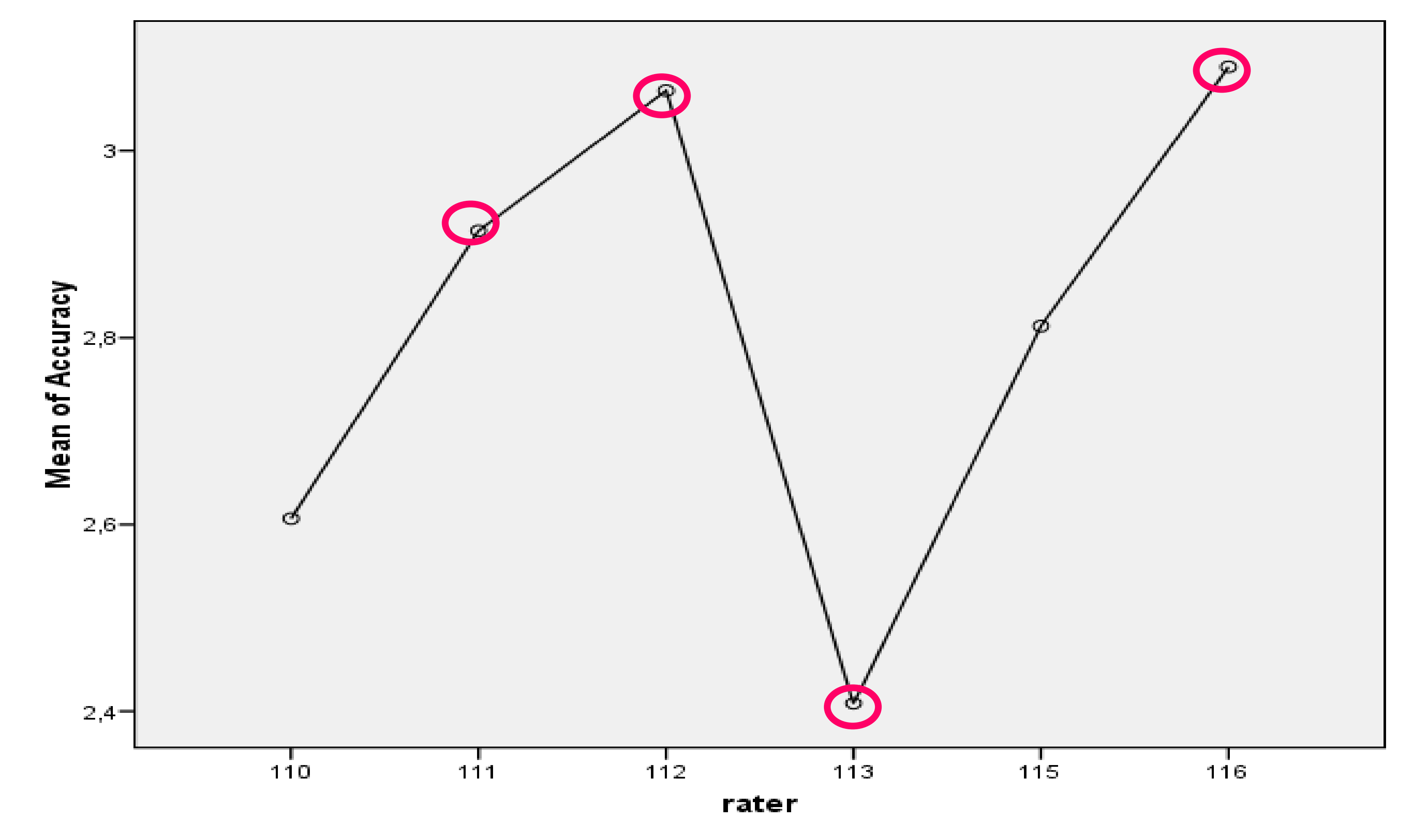
RQ1a: The degree of consistency between raters of the German jury (inter-rater reliability): Pronunciation



RQ1a: The degree of consistency between raters of the German jury (inter-rater reliability): Range



RQ1a: The degree of consistency between raters of the German jury (inter-rater reliability): Accuracy



RQ2a. What is the relation between a rater´s level ratings and verbal comments? (intra-rater reliability)

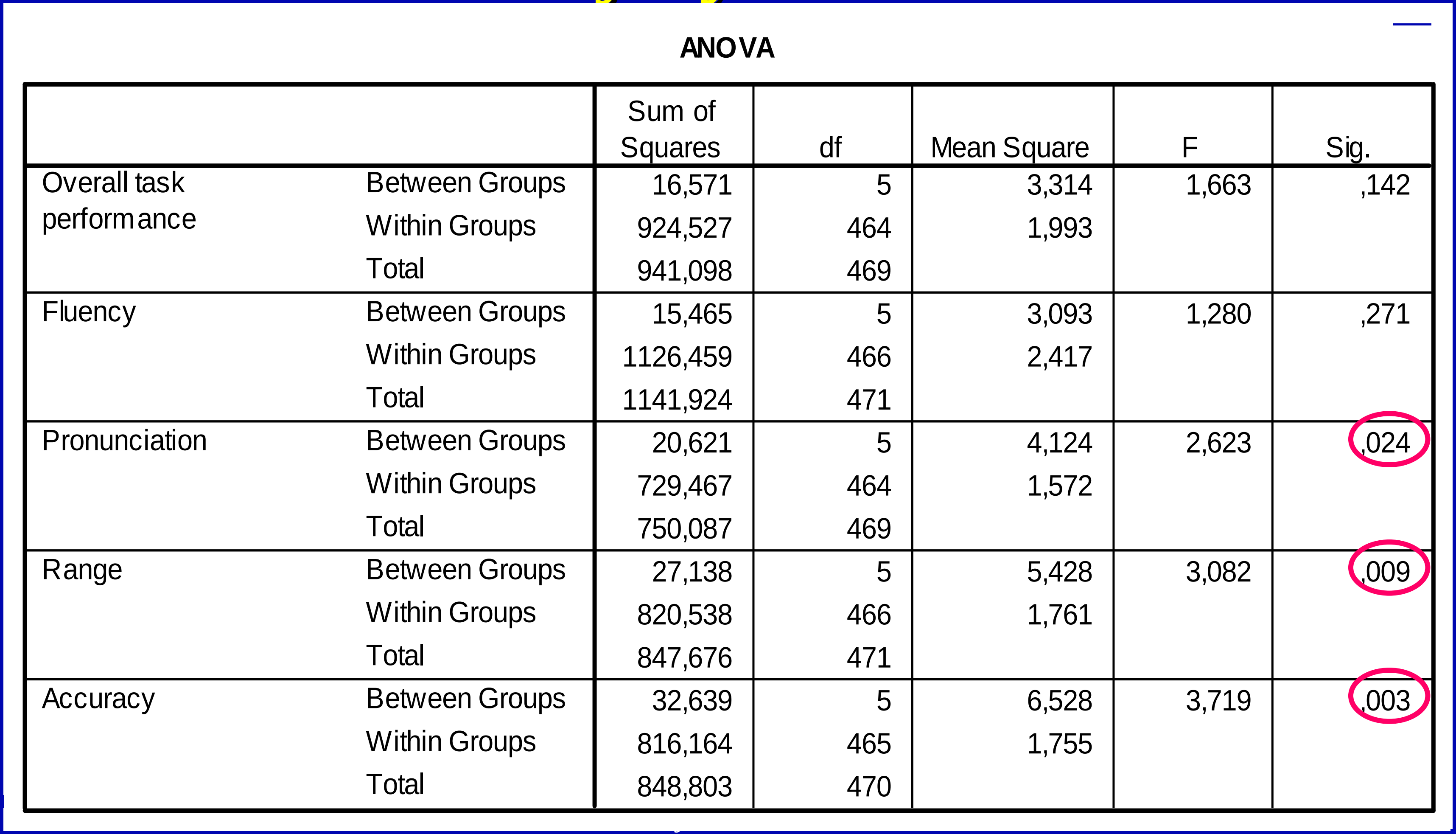
Rater G113 on Fluency: perceived problems with assessing the impact of pauses

Rater G112 on Pronunciation: pays recursive attention to issues of pronunciation

Rater G113 on Range: comments on limitedness of range

Rater G113 on Accuracy: comments on word order and grammatical difficulty

RQ1b. Differences between raters of the German jury on dimensions



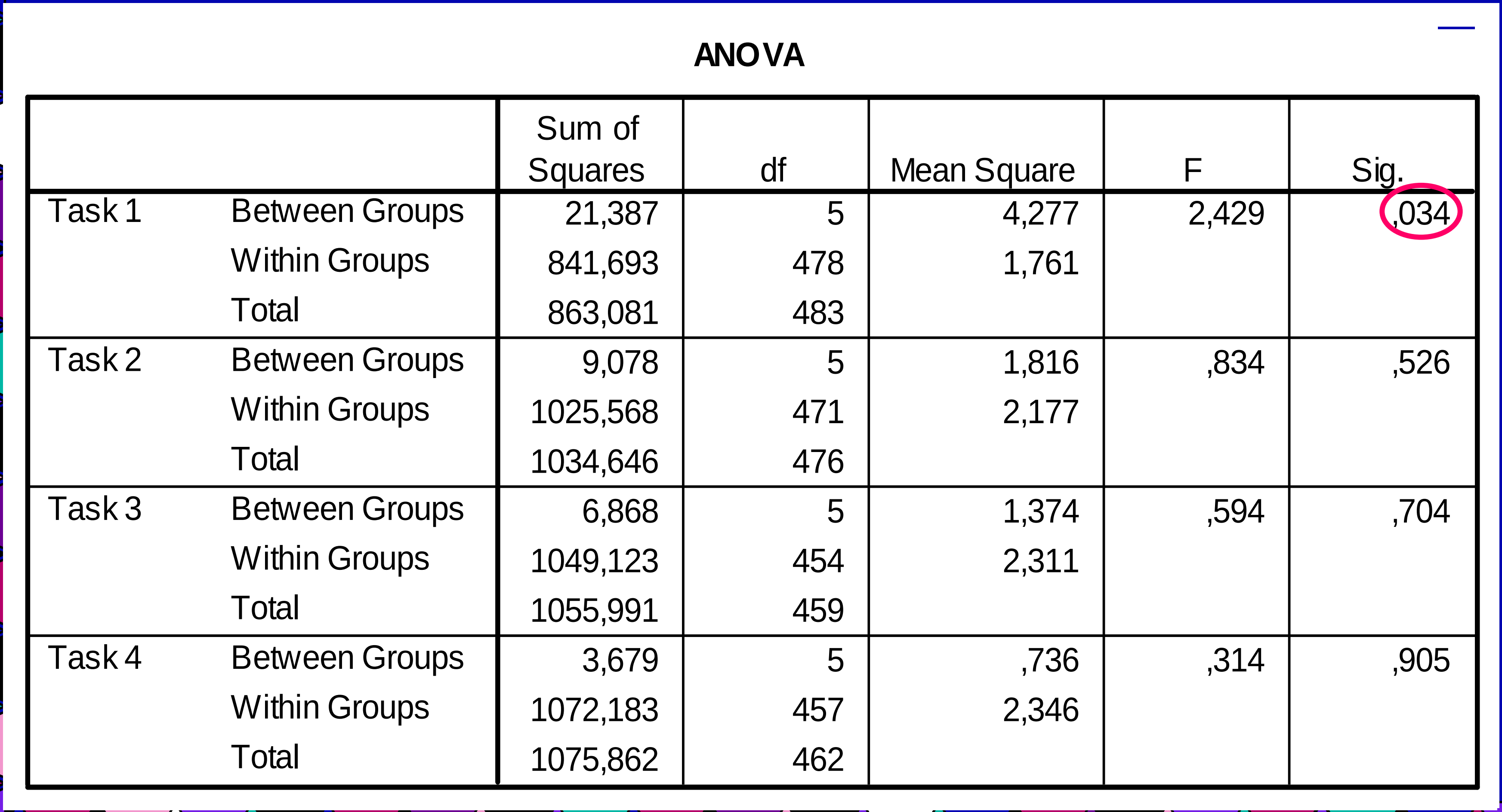
ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Overall task performance	Between Groups	16,571	5	3,314	1,663	,142
	Within Groups	924,527	464	1,993		
	Total	941,098	469			
Fluency	Between Groups	15,465	5	3,093	1,280	,271
	Within Groups	1126,459	466	2,417		
	Total	1141,924	471			
Pronunciation	Between Groups	20,621	5	4,124	2,623	,024
	Within Groups	729,467	464	1,572		
	Total	750,087	469			
Range	Between Groups	27,138	5	5,428	3,082	,009
	Within Groups	820,538	466	1,761		
	Total	847,676	471			
Accuracy	Between Groups	32,639	5	6,528	3,719	,003
	Within Groups	816,164	465	1,755		
	Total	848,803	470			

2b. What is the relation between jury ratings and content of subsequent discussion (inter-rater reliability)

- Quantitative indicators show there is a need of discussion on pronunciation, range and accuracy
- However, the most prominent theme sequences dealt with pronunciation and fluency (most extensive and elaborated, recurrent)
- Accuracy was adressed in a few statements
- Linguistic range appeared in a few separate cases

RQ1b. Differences between raters of the German jury on tasks



ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Task 1	Between Groups	21,387	5	4,277	2,429	,034
	Within Groups	841,693	478	1,761		
	Total	863,081	483			
Task 2	Between Groups	9,078	5	1,816	,834	,526
	Within Groups	1025,568	471	2,177		
	Total	1034,646	476			
Task 3	Between Groups	6,868	5	1,374	,594	,704
	Within Groups	1049,123	454	2,311		
	Total	1055,991	459			
Task 4	Between Groups	3,679	5	,736	,314	,905
	Within Groups	1072,183	457	2,346		
	Total	1075,862	462			

2a. What is the relation between a rater´s level ratings and verbal comments? (intra-rater reliability)

Rater G112 on monologic Task 1: the level of performance was at its best in the beginning of the test

No other task specific comments

RQ3: 3. Themes introduced by raters when motivating their assignments

- 1. Scale
  - General comments on the scale *(The descriptors don´t fit all tasks)*
  - Comments on subscales: overall task performance, fluency, pronunciation, range, accuracy *(How can we define fluency? Accuracy really plays a big role (for instance: for you = auf dich)*
  - Other scale-related comments *(3-4 level descriptions very close to each other, hard to keep apart)*

RQ3: 3. Themes introduced by raters when motivating their assignments

- 2. Factors that motivate the overall judgement:
  - Student-related factors *(They were persistent, so they do deserve A1.3; The pupils were not used to talk at all; Student E didn´t have as many breakdowns as student S (comparison); No more preparation time, the students were bored and wished to leave)*
  - Rater-related factors *(luckily I´m completely unaware of the target level)*

RQ3: 3. Themes introduced by raters when motivating their assignments

- Situation-related factors (*I cannot trust their ability to cope with the situations in real life without instructions; In a paired speaking test it may happen that if one student cannot say what he is supposed to say, also the other one misses the point; Dropping the verb needed deprived the test partner his chance to react*)
- Other factors (*How much does it affect the overall performance if they leave out something from the task instruction)*

RQ3: 3. Themes introduced by raters when motivating their assignments

- 3. Task features
  - Level (*I wonder if the tasks should be longer and more extensive so that students could develop their own ideas; The task restrict student ability, no way to display high level ability)*
  - Structure (*The task layout makes the students to stick to the paper; The warm-up sequence is supposed to loosen the tongue, but in fact, many students got even more strained)*

RQ3: 3. Themes introduced by raters when motivating their assignments

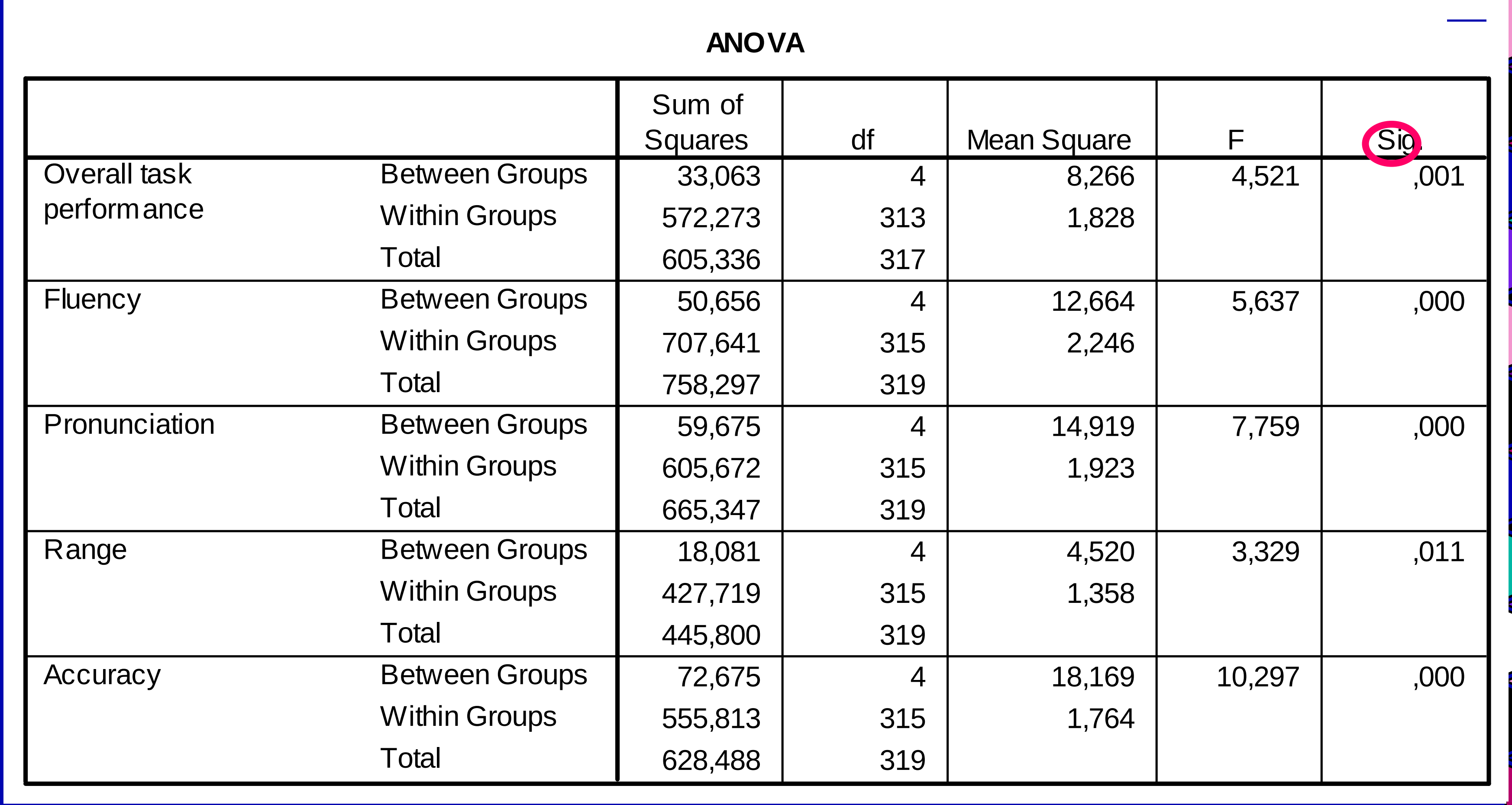
- Themes (*Would it be nicer if the students could choose their topic of discussion; In regular teaching you first learn things in language for a long time, only later on you start speaking little by litte; so you are never told to discuss films just like that…)*

RQ4: Features and patterns of interaction in the discussions of the German jury

- Back channelling, co-contructed turns
- No longer pauses
- Thematic sequences varied in length
- The regular structure of 3 turns (introduction – respose – confirmation)
- Certain topics were revisited by the same raters recurrently
- Features of free discourse: free choice and introduction of of topics, humor
- The discussions get shorter with the passage of time

Tentative results of the English jury:

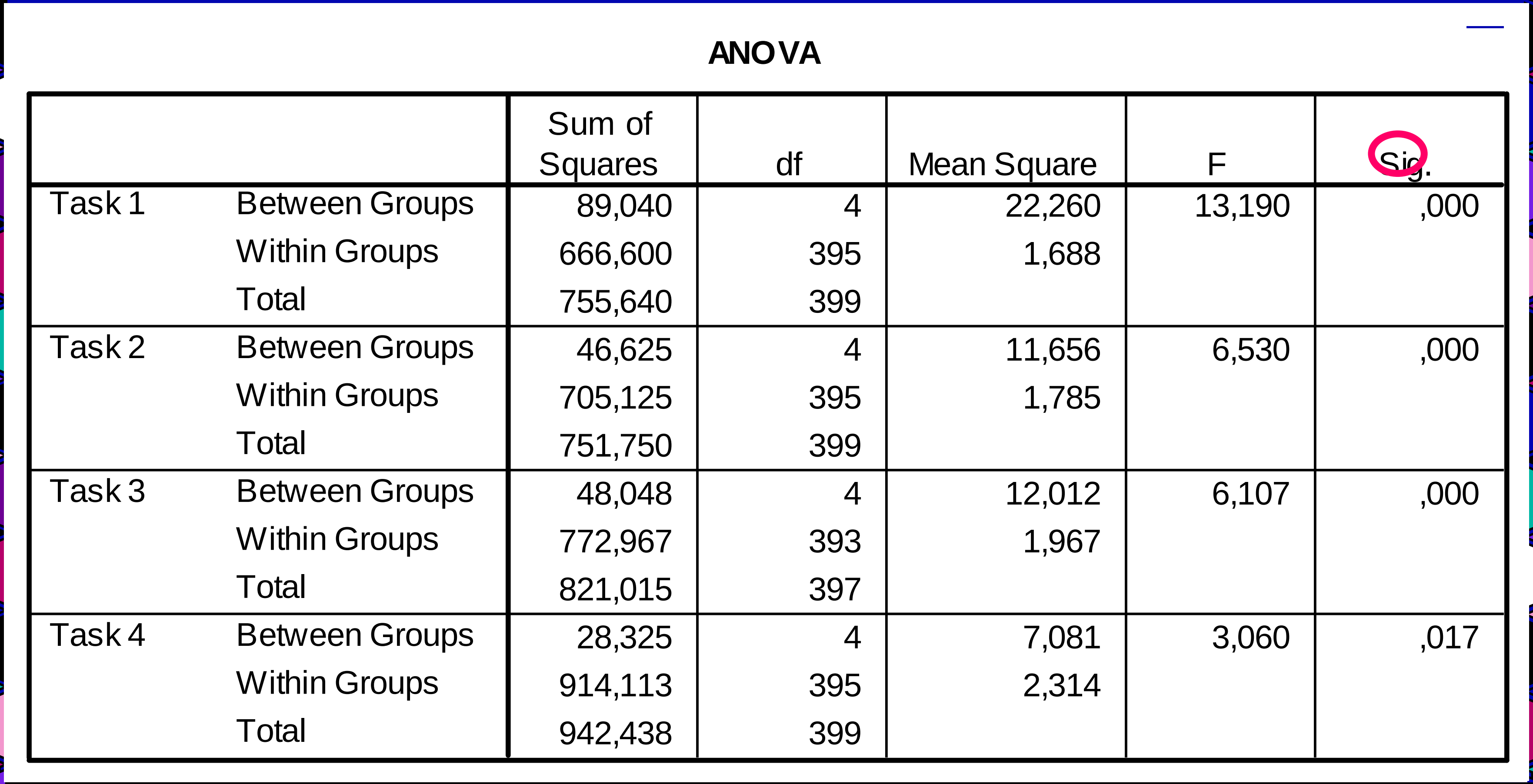
RQ1b. Differences between raters on dimensions



ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Overall task performance	Between Groups	33,063	4	8,266	4,521	,001
	Within Groups	572,273	313	1,828		
	Total	605,336	317			
Fluency	Between Groups	50,656	4	12,664	5,637	,000
	Within Groups	707,641	315	2,246		
	Total	758,297	319			
Pronunciation	Between Groups	59,675	4	14,919	7,759	,000
	Within Groups	605,672	315	1,923		
	Total	665,347	319			
Range	Between Groups	18,081	4	4,520	3,329	,011
	Within Groups	427,719	315	1,358		
	Total	445,800	319			
Accuracy	Between Groups	72,675	4	18,169	10,297	,000
	Within Groups	555,813	315	1,764		
	Total	628,488	319			

Tentative results of the English jury: RQ1b. Differences between raters on tasks



ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Task 1	Between Groups	89,040	4	22,260	13,190	,000
	Within Groups	666,600	395	1,688		
	Total	755,640	399			
Task 2	Between Groups	46,625	4	11,656	6,530	,000
	Within Groups	705,125	395	1,785		
	Total	751,750	399			
Task 3	Between Groups	48,048	4	12,012	6,107	,000
	Within Groups	772,967	393	1,967		
	Total	821,015	397			
Task 4	Between Groups	28,325	4	7,081	3,060	,017
	Within Groups	914,113	395	2,314		
	Total	942,438	399			

Tentative results of the English jury compared to German data:

- Suggest that there is larger variance between raters and lower reliability
- Shorter discussions
- Less elaborated sequences on themes, typically single statements

Results →Relevance, utility, sufficiency

Rebuttals against the claim in regard to

- Certain dimensions of the rating scale: pronunciation, range and accuracy (utility)
- Monologic task (relevance)
- Structure of dialogic tasks (relevance, utility, sufficiency)

Conclusions & recommendations

- Rater training
- Task development
- Scale development

KIITOS

Raili.hilden@helsinki.fi
Marja.k.martikainen@helsinki.fi