Technology-Assisted Scoring of Short-Answer Items for Listening Comprehension: A Clustering Approach

Leska Schwarz, Christian Gold

EALTA 2019, Dublin

TestDaF
Test Deutsch als Fremdsprache

g.a.s.t.

LANGUAGE TECHNOLOGY LAB

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

Scoring of Listening Comprehension Tasks

Sie hören einen Ausschnitt aus einer Vorlesung im Fach Neurowissenschaften zum Thema „Gesichtserkennung“. Ergänzen Sie beim Hören die Gliederungspunkte in Stichpunkten. /

You will hear part of a lecture on the topic of “face recognition”. While listening, complete the bullet points.

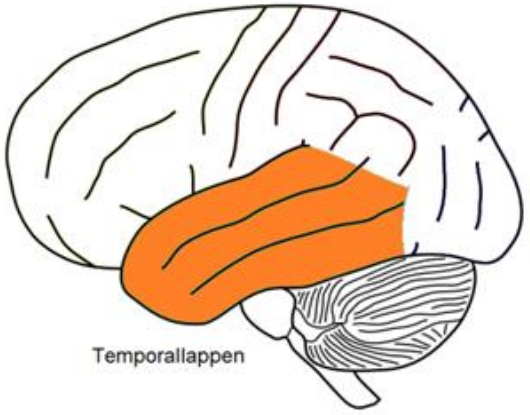
Item

Merkmal von Gesichtszellen /

feature of face recognition cells:

sind besonders aktiv beim Betrachten von Gesichtern /

are active when seeing a face



Motivation

problem of short-answer questions:
variety of answers → decision whether response is correct or not

✓ sind besonders aktiv beim Betrachten von Gesichtern

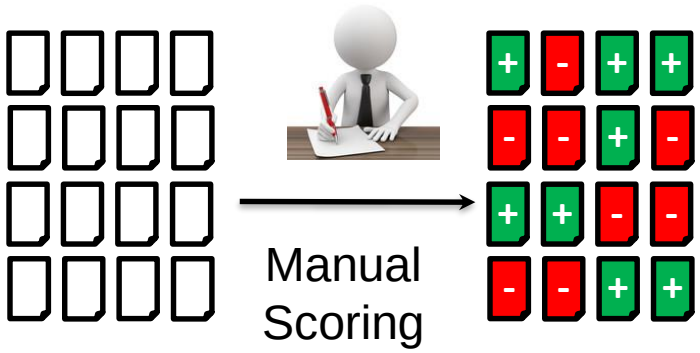
X Sind sehr aktiv

X aktive Erkennung von Gesicht

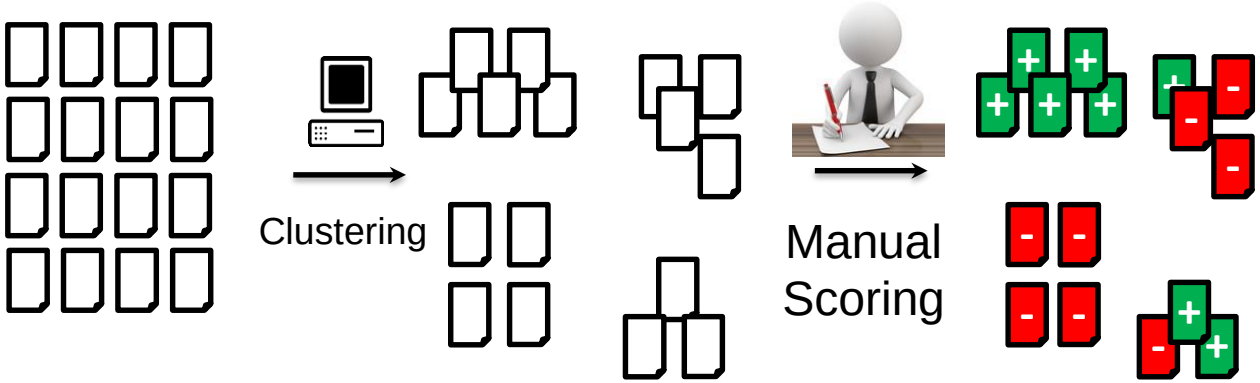
? sind besonders aktiv bei Betrachten von Gesichter

Clustering for Easier Scoring

Traditional Scoring



Cluster-Based Scoring



Clustering Examples

- *Hirnbereiche in Temporallapen* ✓
- *Temporallappen* ✗
- *Temporallappen ist aktiv* ✗
- *Trmporallappen* ✗

- *affen* ✗
- *aktiv* ✗
- *durch Tomogeraphy* ✗
- *miliarden* ✗
- *orange gefärbt* ✗

- *Besonders aktiv beim Anblick der Gesichter.* ✓
- *beim Betrachten der Gesichter aktiv* ✓
- *aktive bei der Beachtung aud gesicht* ✓
- *sie sind beim Einblick bekannter Gegenständen aktiv* ✗

Clustering Setup:

- 10 clusters per item
- k-Means clustering based on token and character n-grams
- alphabetically sorted

Research Hypotheses

Hypotheses:

- Scoring clustered items is **faster** than scoring unclustered items.
- Scoring clustered items yields a **higher inter-rater agreement** between raters.

Data

- TestDaF: test for German as a foreign language for university admission
- transition from paper-based to computer-based testing

- 3 test sets
- 620 test-takers
- overall 7530 responses

test set	
	items
Task 1	5
Task 2	4
Task 3	4

Scoring Guidelines / Key

Aufgabe 5 (54)

Item 54.1: Merkmal von Gesichtszellen

Richtig:

- besonders aktiv bei der Betrachtung von Gesichtern
- aktiv, wenn wir ein Gesicht sehen
- befinden sich im Temporallappen
- erkennen / unterscheiden Gesichter

Falsch:

- sind sehr aktiv [zu allgemein]
- Temporallappen [zu verkürzt]

Wichtig bei diesem Item:

- Falsch gebildete Plural-Formen von „Gesicht“ werden als richtig akzeptiert.
- Auch wenn sie eigentlich kein *Merkmal* der Gesichtszellen ist, wird ihre *Hauptaufgabe*, also das Betrachten / Erkennen / Unterscheiden von Gesichtern, in diesem Item als richtig bewertet.

definitely right

definitely wrong

explanation

Study Design

Study 1: 4 human raters, ~ 200 test-takers

	first half of items	**second half of items**
rater A + B	clustered	random
rater C + D	random	clustered

Study 2: 6 human raters, 2 sets with ~ 200 test-takers each

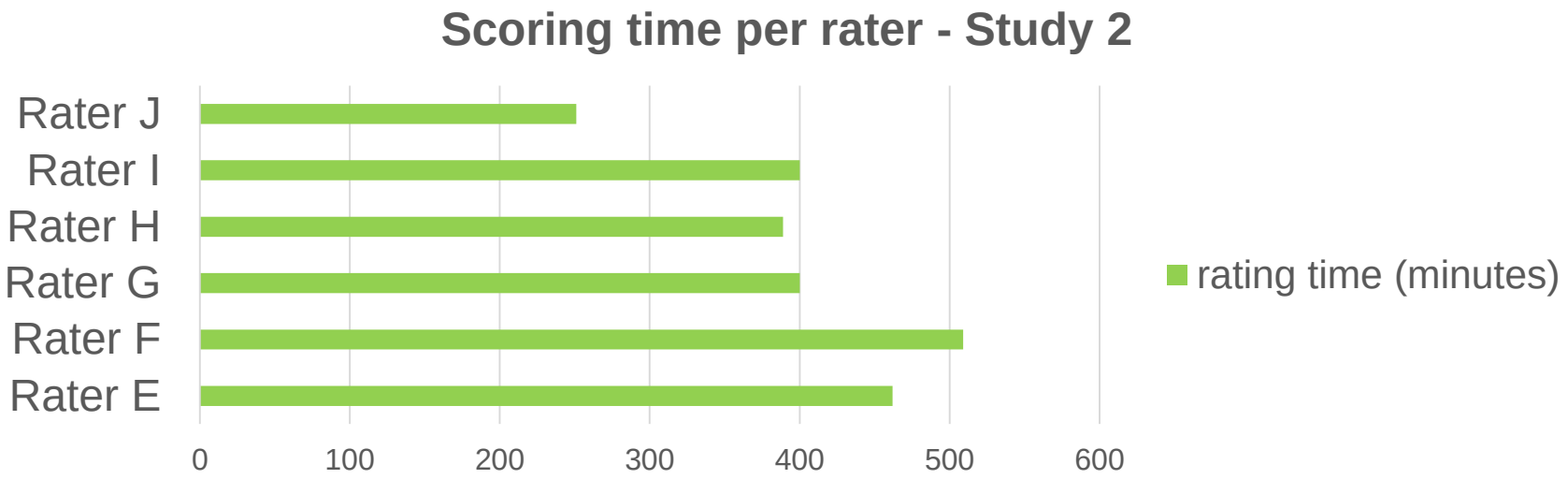
	first third of items	**2nd third of items**	**3rd third of items**
rater E + F	clustered	random	alphabetic
rater G + H	alphabetic	clustered	random
rater I + J	random	alphabetic	clustered

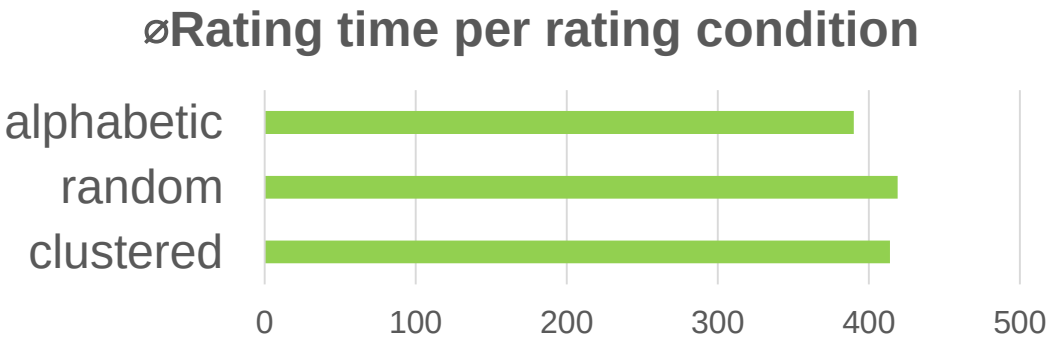
Rating Times

Research Question: Are clustered answers rated faster?

Rating Times

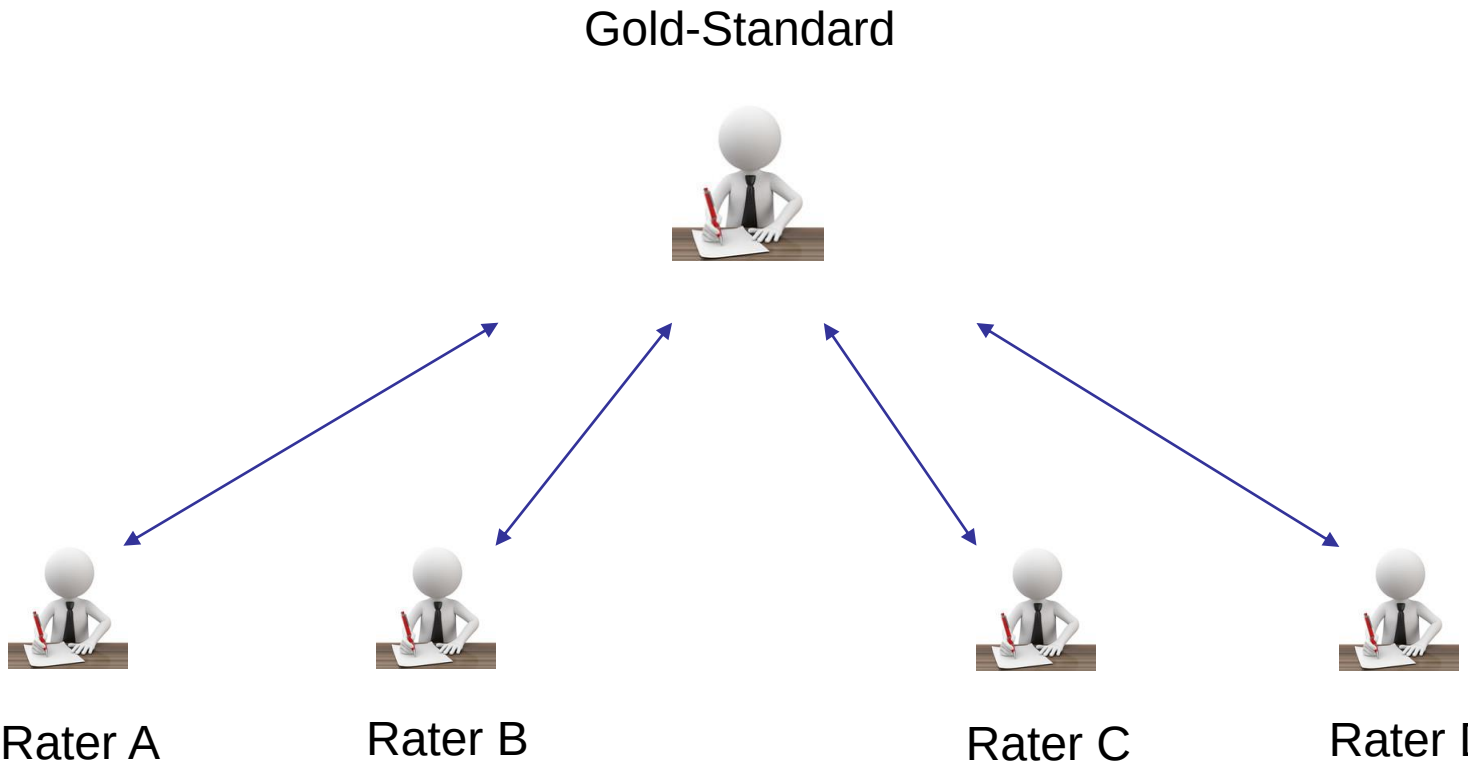
Research Question: Are clustered answers rated faster?





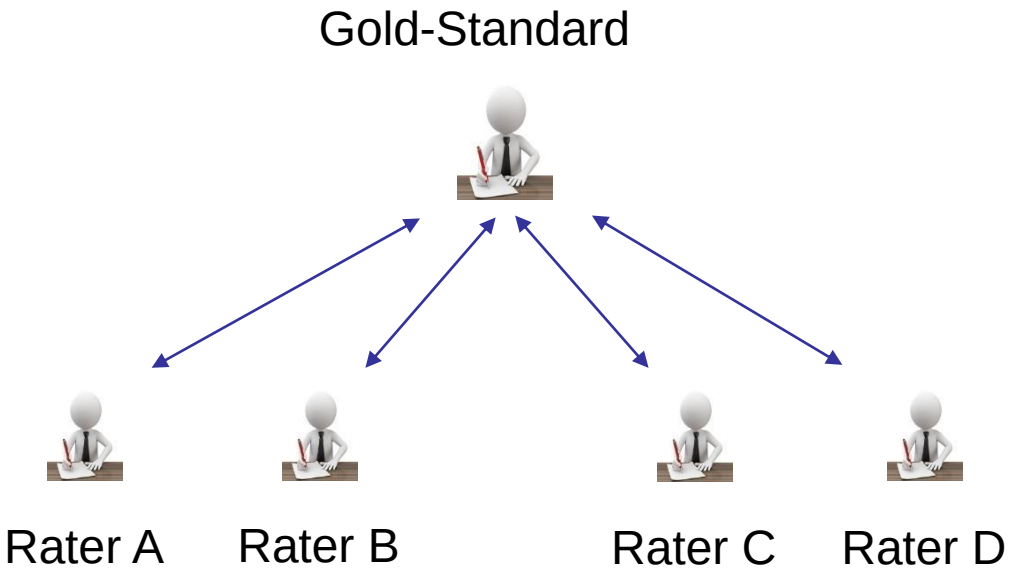
Inter-Rater-Agreement

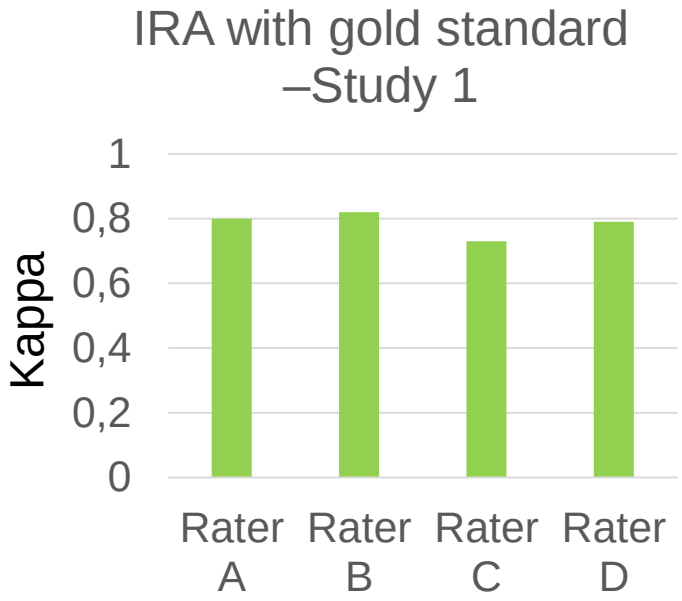
Research Question: Are clustered answers rated more consistently?



Inter-Rater-Agreement

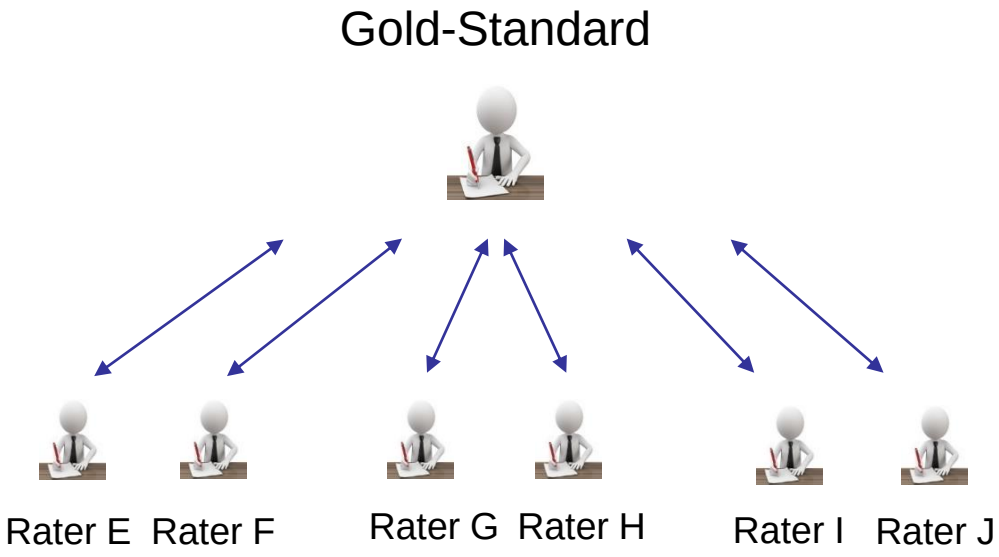
Research Question: Are clustered answers rated more consistently?

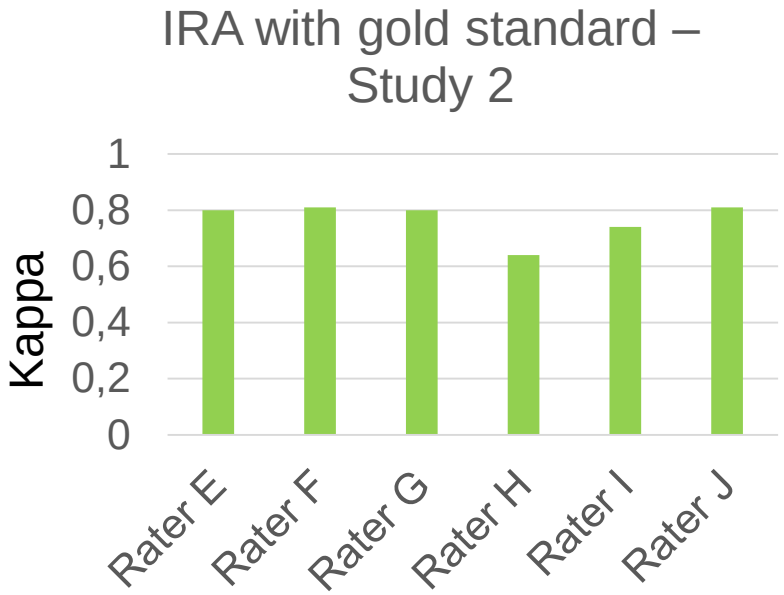




Inter-Rater-Agreement

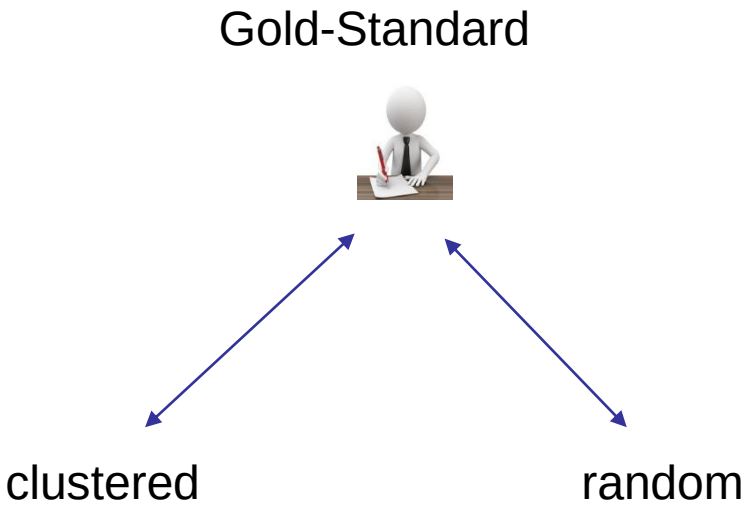
Research Question: Are clustered answers rated more consistently?





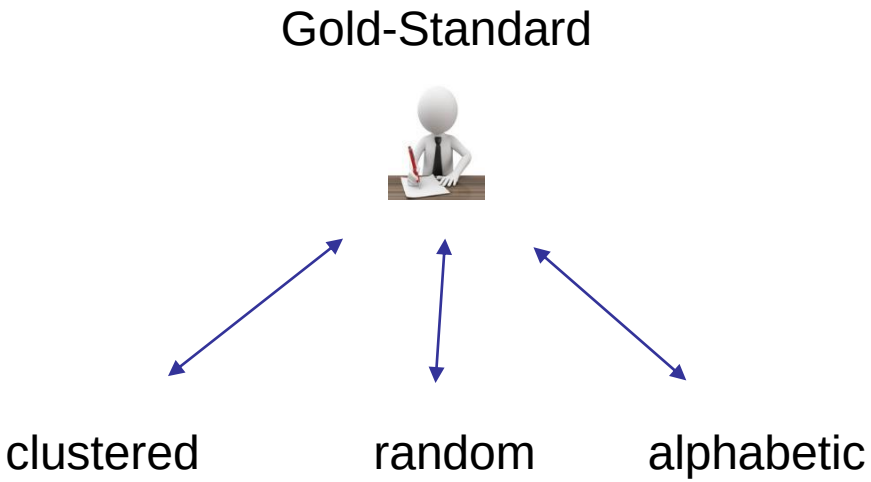
Inter-Rater-Agreement

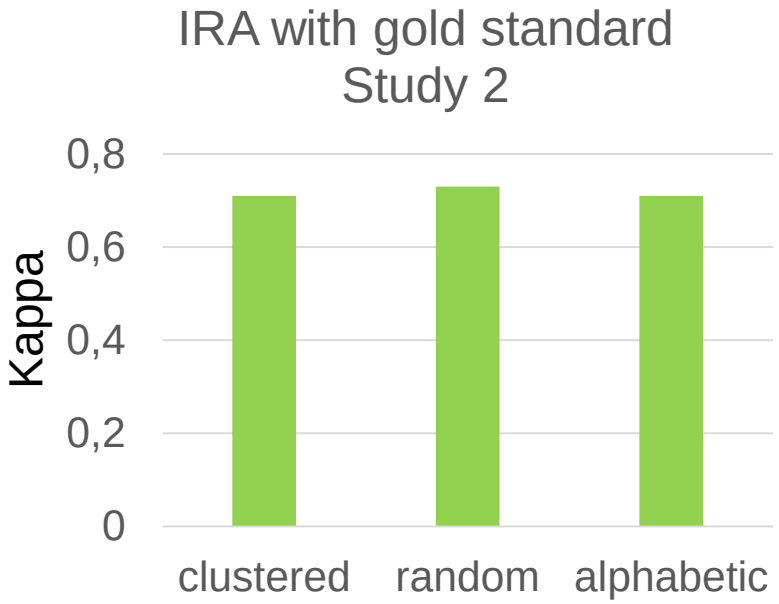
Research Question: Are clustered answers rated more consistently?



Inter-Rater-Agreement

Research Question: Are clustered answers rated more consistently?





Qualitative Feedback

Overall very positive:

- time saving
- extremely helpful
- focus on one scoring aspect only, not so much going back and forth between aspects
- scoring seems to be more structured

Conclusions and Outlook

- Clustered items are perceived as helpful, but it seems difficult to prove that quantitatively.

- Different setup needed?

- Ideas and things to consider
  - longitudinal approach
  - differences between items
  - rater effects
  - think-aloud protocols
  - pre-scored responses
  - …

Thank you!

Leska Schwarz leska.schwarz@testdaf.de
Christian Gold christian.gold@uni-due.de