EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

LEAD
Graduate School

Computer-assisted Scoring of Short Responses: The Efficiency of a Clustering-based Approach in a Real-life Task

Magdalena Wolska[1], Andrea Horbach[2], and Alexis Palmer[2]

[1]Eberhard Karls Universität Tübingen
[2]Universität des Saarlandes

Constructed response items

test-taker required to *produce* a response, rather than select one from a predefined set

measure complex skills

can be applied to various tasks (describe, summarise, formalize, ...)

Constructed response items

- scoring process requires judgment
- time-consuming
- higher cost of testing, longer time need to calculate and deliver scores

Constructed response items
- scoring process requires judgment
- time-consuming
- higher cost of testing, longer time need to calculate and deliver scores

„In the interests of economy [...] and to take advantage of the capacity for automated coding that the electronic medium offers, **for the electronic reading assessment a higher proportion of items requiring no coder judgment has been included**"
PISA **2009** Assessment Framework

„**The labor required to score CR items is a major assessment cost.** A variety of systems have been or are being developed and placed in service to automatically score student essay and other CR items using AI engines. Based on ASG's research, today these systems cost between $.50 and $3 per response with the bulk of the pricing by vendors at the higher end of the range. **It is assumed that as time passes and systems continue to mature, pricing should become more affordable.**"
Stanford Center for Opportunity Policy in Education **2010** Report „The Cost of New Higher Quality Assessments"

Constructed response items
- scoring process requires judgment
- time-consuming
- higher cost of testing, longer time need to calculate and deliver scores

„In the interests of economy [...] and to take advantage of the capacity for automated coding that the electronic medium offers, **for the electronic reading assessment a higher proportion of items requiring no coder judgment has been included**"
PISA **2009** Assessment Framework

„**The labor required to score CR items is a major assessment cost.** A variety of systems have been or are being developed and placed in service to automatically score student essay and other CR items using AI engines. Based on ASG's research, today these systems cost between $.50 and $3 per response with the bulk of the pricing by vendors at the higher end of the range. **It is assumed that as time passes and systems continue to mature, pricing should become more affordable.**"
Stanford Center for Opportunity Policy in Education **2010** Report „The Cost of New Higher Quality Assessments"

unclear whether selected-response item can provide the same information about the test-taker group differences in performance on selected vs. constructed response items have been observed

Constructed response items

- most of to-date research has focused on automated scoring
  - Callear et al.; Bachman et al.; Mitchell et al.
  - C-rater
  - Sukkarieh&Pulman
  - Bailey&Meurers
  - Mohler&Mihalcea
  - Meurers et al., Ziai et al. Ott et al.
  - UKP-BIU
  - Kaggle challenge
  - SemEval task

Constructed response items

- most of to-date research has focused on automated scoring
  - Callear et al.; Bachman et al.; Mitchell et al.
  - C-rater
  - Sukkarieh&Pulman
  - Bailey&Meurers
  - Mohler&Mihalcea
  - Meurers et al., Ziai et al. Ott et al.
  - UKP-BIU
  - Kaggle challenge
  - SemEval task

- alternative: computer-***assisted*** scoring
  - Basu et al., 2013 (Microsoft): „Powergrading: A Clustering Approach to Amplify Human Effort for Short Answer Grading”

  - our work: developped in parallel, along the same idea

Premise

„similar” responses express the same content,
thus are likely to receive the same score

Premise

„similar” responses express the same content,
thus are likely to receive the same score

Questions

By what order of magnitude can the number of responses
to score be reduced while maintaining acceptable scoring accuracy?

Does grouping similar responses make manual scoring faster?

Premise

„similar" responses express the same content,
thus are likely to receive the same score

Questions

By what order of magnitude can the number of responses
to score be reduced while maintaining acceptable scoring accuracy?
→ accuracy vs. rater's workload
intrinsic evaluation via simulation

LREC-14

Does grouping similar responses make manual scoring faster?
→ time-on-task
task-based evaluation in a real-life setting

This work

Outline

Data

Clustering: features, method, parameter

Experiment: Timing the scoring task

Conclusions and further work

Data

Data

Placement tests for DaF courses at Saarland University

Tests administered via online platform

Listening comprehension (LC):

- short constructed response items
- testlet: 3 audio prompts of increasing difficulty
- about 24 questions per testlet
- test-takers can read questions before/during listening
- audio can be played twice for most of the items
- test-takers allowed to take notes on paper

- most questions scored on 1-point scale
- partial credits at .5-points
- scoring done by teachers

Data

Woher kommt Norma?

Where does Norma come from?

Angloa
Angol
angola
Angola
Ängola
Angolia
Ängolij
Angolla
angorla
Angula
aus Anggola
aus angloa
aus Angola
Aus Angola
Aus Angola.

Aus Angora
Aus Andorra
aus Engola
aus ingol
England
Engola
Engula
Berlin

Norma kommt aus Angola
Norma komms aus Angola
Norma kommst aus Angola.
Norma kommst aus Angola.
Norma kommt auf Angola
Norma kommt aus Angola.
Norma kommt aus Angolla.
Norma kommt von Angola
Norma komt aus Angola
Momma kommt aus Angola.

Er kommt aus Angola
Sie komme aus Angola
sie kommt aus Angola
Sie kommt aus Angola
Sie kommt aus Angola.
sie kommt aus Angolla
Sie kommt aus Angora.
Aus Angola kommt sie.
Norma

Data

Warum sieht Herr Wienert jetzt nicht mehr so viel fern?

Why doesn't Mr Wienert watch so much TV anymore?

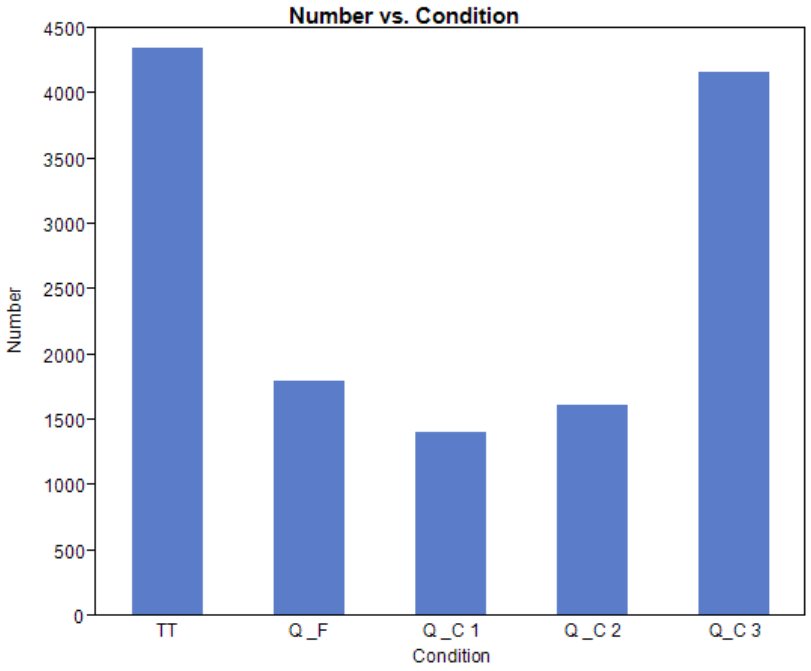
er is aus
keine zeite
Besuche mehr Kultur.
aus zum Theatre, Kino
kino, theatre, Frankfurt
Er wohnt zusamen mit Marita
Abends ins Kino oder ins Theater
besser ins theater oder ins kino gehen
das kulturelles Angebot ist meh interessant, Kino und Theater
Er geht am Abnd lieber ins kino, seit er mit Marita zusammen ist .
er hat eine Freundin und bevorzugt ins Kino oder Theater zu gehen
Da er in Frankfurt woht, es es ihn leichter ins Kino oder Theater geht.
Er gehe ins Kino oder ins Teather mit seine Frau und sie haben kein Zeit.
davor hat er mehr Zeit und jetzt hat er verschiedenen Kulturelles angeboten
am Abend ins Kino,Theater geht,es gibt viele kullturelle Angebote in Frankfurt
er geht jetzt viel ins kino, in theater: es gibt eine reiche Kultural Angebot in Frankfurt
Da er mit jemandem wohnt, sieht er jetzt nicht so viel fern, weil sie oft ins Theater oder ins Kino gehen.
Er geht ins Kino und Theater. Herr Wienert lebt in München und es gibt reiches kulturelles Angebot dort.
Als er jetzt wohnt mit Rita zusammen, und möchte gern in Kino oder Theater gehen; wo Herr Wienert lebt, gibt es viele kulturelle Angebote.
Denn er wohnt zusammen mit seiner Freundin und sie gehen ins Kino und ins Theater zusammen. Deshalb hat er nicht so viel Zeit, um fernzusehen.
Weil er hat jetzt viel Arbeit zu tun, ausserdem jetzt er keine Freundin hat. Er hat viel mehr TV mit seine Alter-Freundin gesehen, vorher sie sind auch ins Kino und Theater zusammen gegangen, usw.

(usw.)

Data

5 placement test rounds (April-October 2013)

response tokens



Clustering the responses

Clustering the responses

Preprocessing

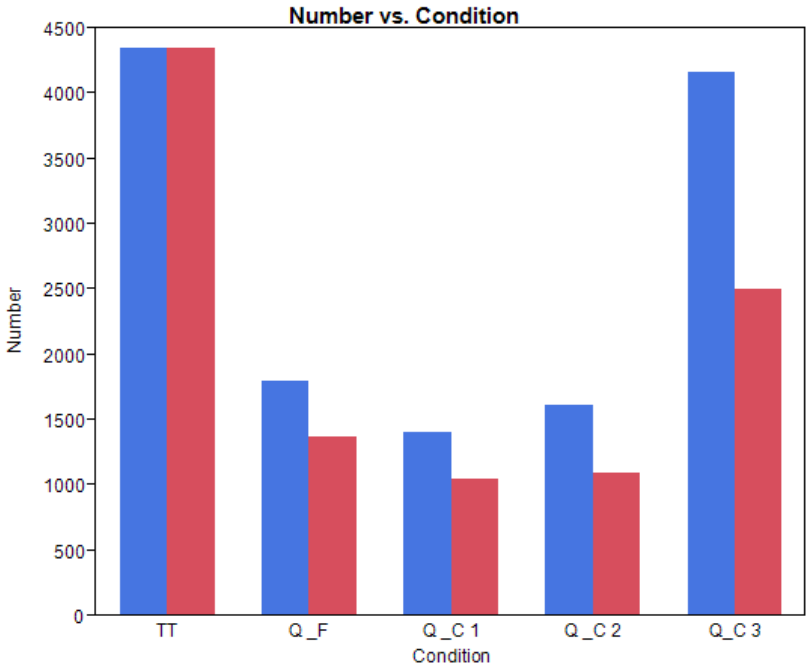
- removing punctuation, lemmatization (TreeTagger), lowercasing
- collapsing token-identical strings to single observation

Clustering the responses

Preprocessing

removing punctuation, lemmatization (TreeTagger), lowercasing

collapsing token-identical strings to single observation



Clustering the responses

Features

- n-grams
- keywords
- question material

Clustering the responses

Features

- n-grams
  - word n-grams {1,2,3}, skip n-grams {2,3}
  - character n-grams before lemmatization {1,2,3,4}
- keywords (**KW**)
  - most relevant concepts from target answers
  - simple weighing by repeating 100 times in vectors
- question material (**QM**)
  - remove question's topic lexemes

Clustering the responses

Clustering setup

- single pass clustering

Clustering the responses

Clustering setup

- single pass clustering
- cosine similarity wrt. centroids of created clusters
  - if greater than Threshold, create new cluster
  - otherwise, include in centroid's cluster

- four models

include KW	exclude QW
include KW	don't exclude QW
don't include KW	exclude QW
don't include KW	don't exclude QW

- thresholds
  - [0.1:0.1:0.9]

Clustering the responses

Experimental conditions

3 + 2 sets of test data (5 tests)

3 scoring conditions (scoring sheets presented in one of 3 modes)

Clustering the responses

Experimental conditions

3 + 2 sets of test data (5 tests)

3 scoring conditions (scoring sheets presented in one of 3 modes)

by Test-Taker **TT**

by Question:

responses ordered **by Frequency:** **Q_F**

responses **Clustered:** **Q_C1, Q_C2, Q_C3**

Clustering the responses

Experimental conditions

3 + 2 sets of test data (5 tests)

3 scoring conditions (scoring sheets presented in one of 3 modes)

by Test-Taker	**TT** as in pen-and-paper setup (familiar)
by Question:	
responses ordered **by Frequency:**	**Q_F** ← baseline
responses **Clustered:**	**Q_C1, Q_C2, Q_C3**

Clustering the responses

Experimental conditions

3 + 2 sets of test data (5 tests)

3 scoring conditions (scoring sheets presented in one of 3 modes)

by Test-Taker — **TT**
as in pen-and-paper setup
(familiar)

by Question:

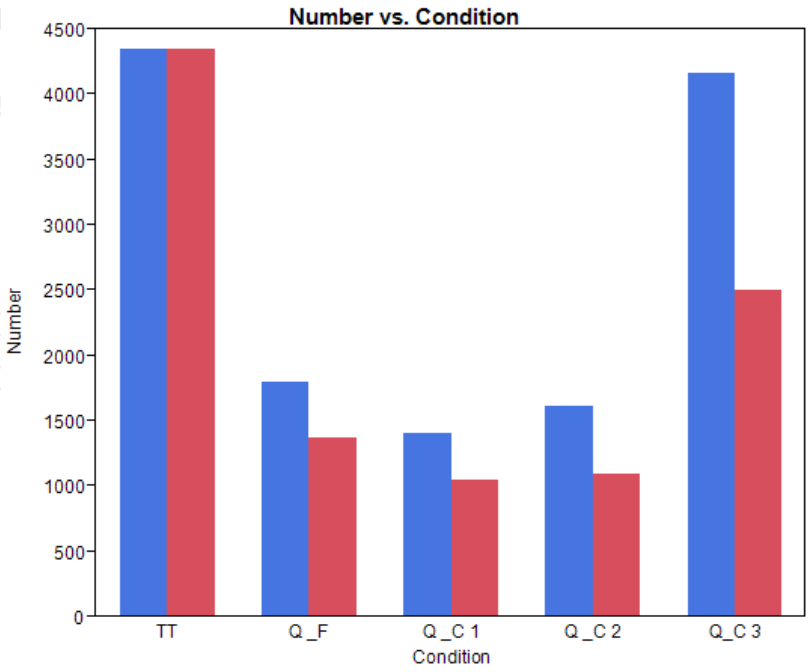
responses ordered **by Frequency:** — **Q_F** ← baseline
responses **Clustered:** — **Q_C1, Q_C2, Q_C3**

response sheet set of responses displayed to score

TT: responses to **all items by one test-taker**
Q: responses to **one item by all test-takers**

Clustering the responses

Experimental conditions



Clustering the responses

Experimental conditions

3 + 2 sets of test data (5 tests)

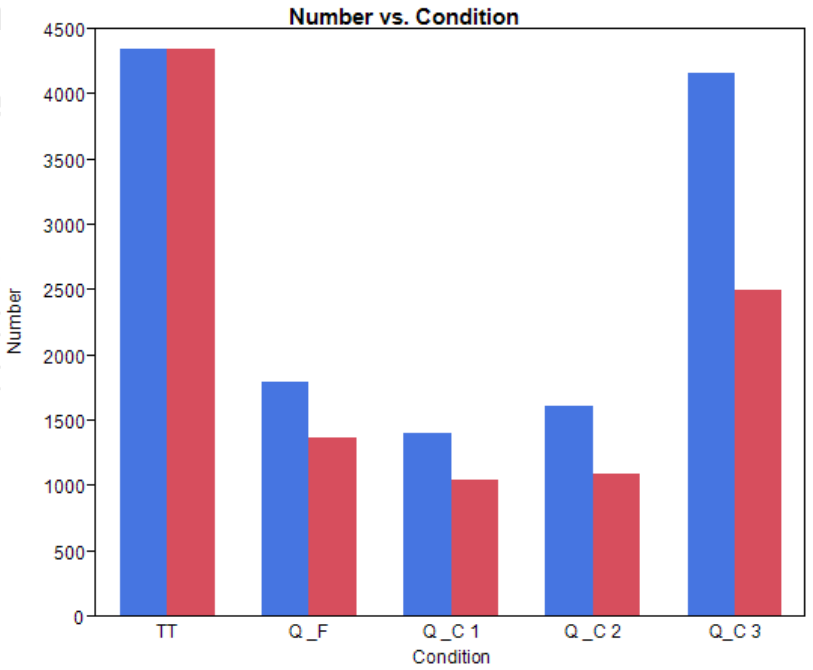
timing data in

by Test-Take

by Question

resp

resp



different LC items at each time

Clustering the responses

Model and parameter selection

Clustering the responses

Model and parameter selection

based on Q_F data (previously scored)

using standard precision-oriented measures: purity and entropy

model: high mean purity / low mean entropy, low variance over all questions and similarity thresholds

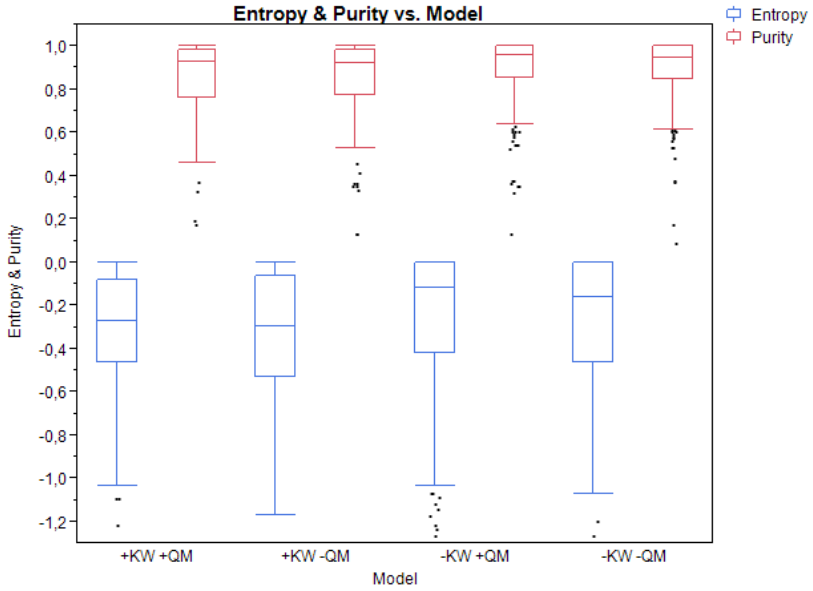
Clustering the responses

Model and parameter selection

based on Q_F data (previously scored)

using standard precision-oriented measures: purity and entropy

model: high mean purity / low mean entropy, low variance over all questions and similarity thresholds



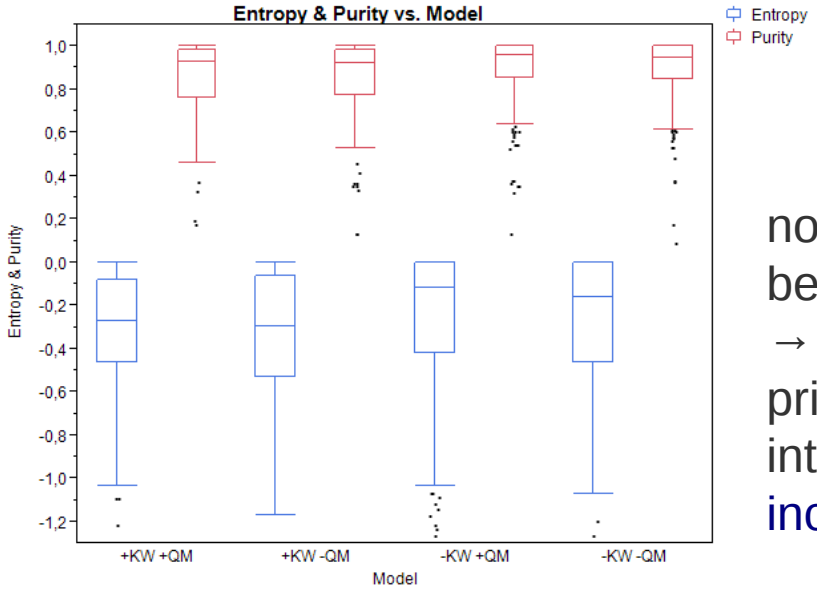
Clustering the responses

Model and parameter selection

based on Q_F data (previously scored)

using standard precision-oriented measures: purity and entropy

model: high mean purity / low mean entropy, low variance over all questions and similarity thresholds



no statistical differences between the models
→ pick model based on prior results and linguistic intuiton:
include KW, exclude QM

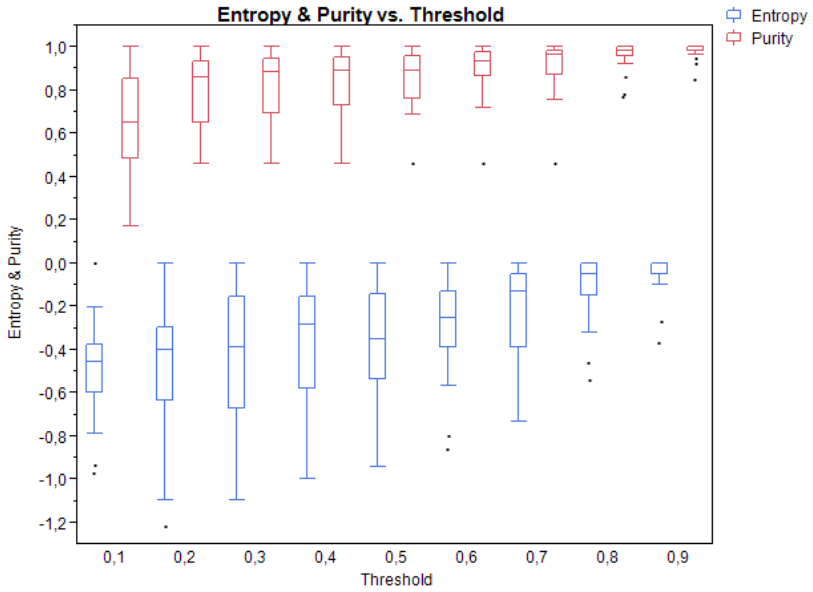
Clustering the responses

Model and parameter selection

based on Q_F data (previously scored)

using standard precision-oriented measures: purity and entropy

threshold for include KW, exclude QM



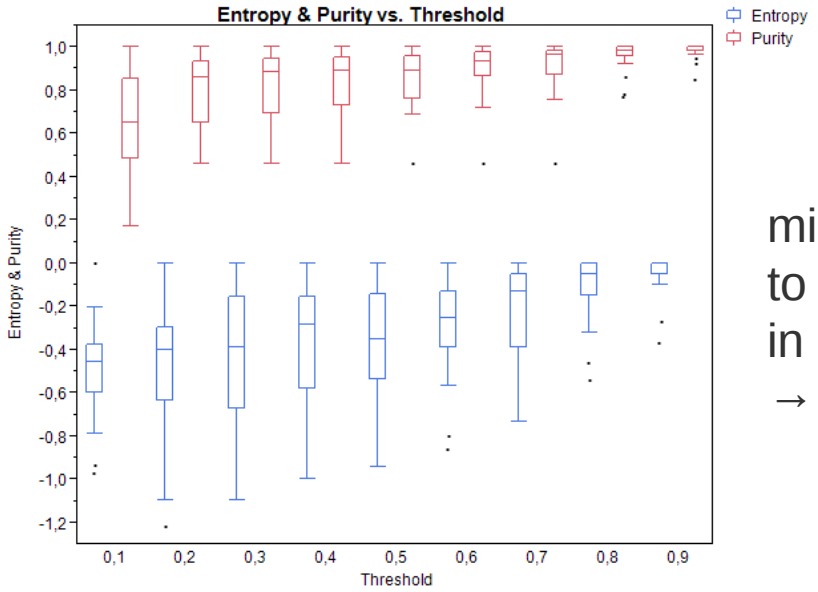
Clustering the responses

Model and parameter selection

based on Q_F data (previously scored)

using standard precision-oriented measures: purity and entropy

threshold for include KW, exclude QM



mid-range
to allow larger differences
in spelling
→ 0.4

Time-on-task

Time-on-task

Is scoring sheets of clustered responses faster than of non-clustered responses?

Time-on-task

Measures

sheet scoring time: from the time sheet opened to submit

per response scoring time: sheet scoring time / No. responses per sheet

Time-on-task

Measures

sheet scoring time: from the time sheet opened to submit
→ selecting dataset for analysis

per response scoring time: sheet scoring time / No. responses per sheet
→ comparisons acorss conditions

Time-on-task

Dataset for analysis

include

response sheets which have been opened once

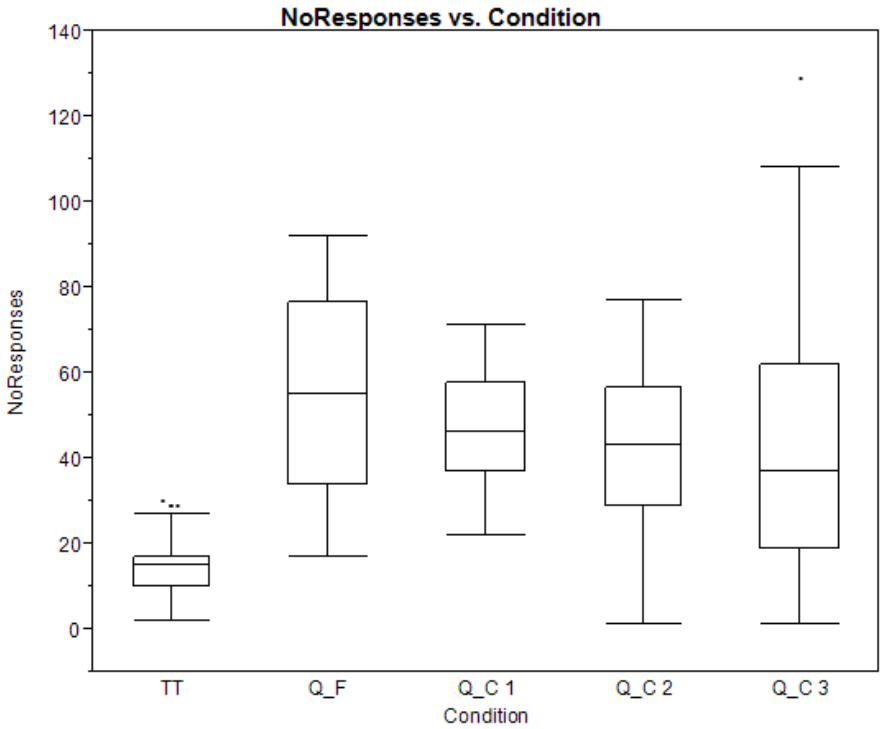
if multiple times, then completely scored the first time opened

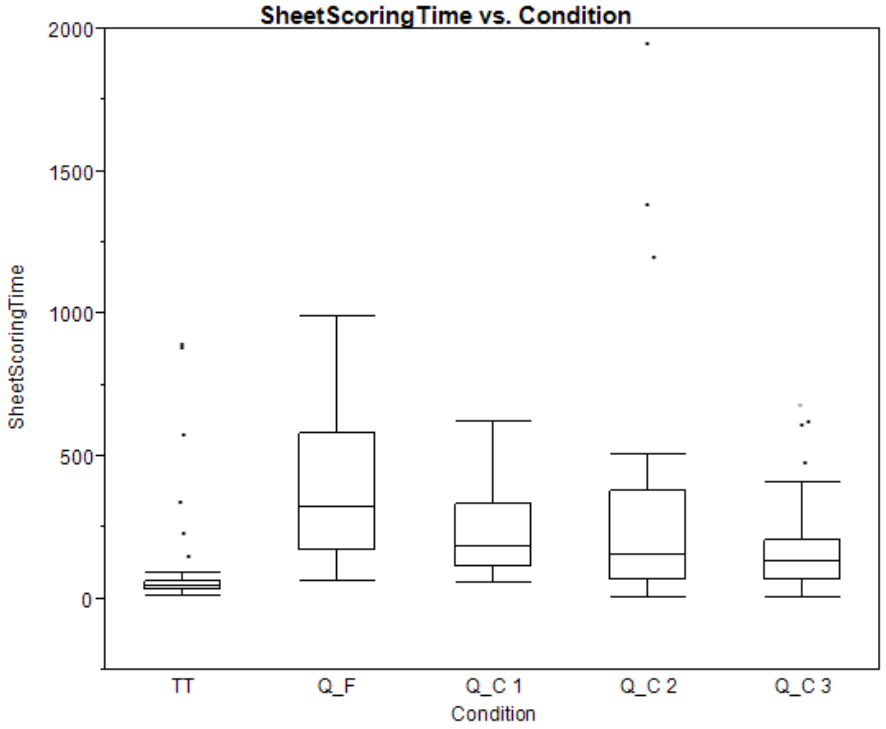
remove response sheets with very small number of responses

remove sheets with unusually long total scoring time

Time-on-task

Dataset for analysis

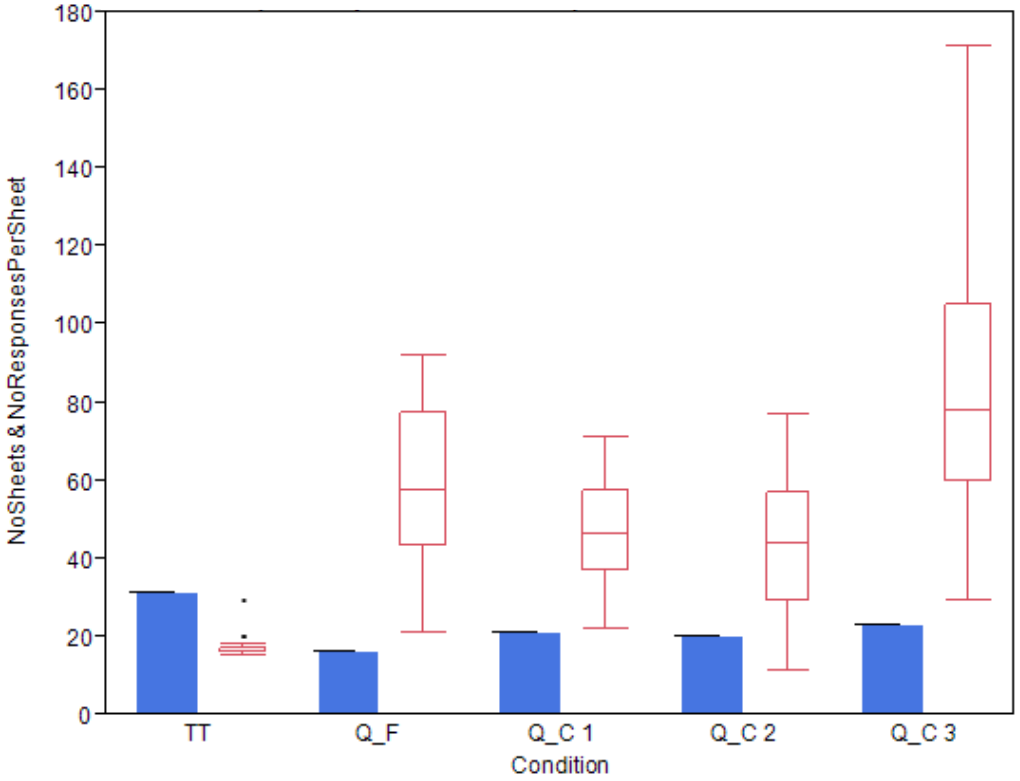




TT: at least 15 responses
(half of max no. responses)

remove all the 11 outliers

Time-on-task


NoSheets & NoResponsesPerSheet
180
160
140
120
100
80
60
40
20
0
TT
Q_F
Q_C 1
Q_C 2
Q_C 3
Condition

Time-on-task

Is scoring sheets of clustered responses faster than of non-clustered responses?

Time-on-task

Is scoring sheets of clustered responses faster than of non-clustered responses?

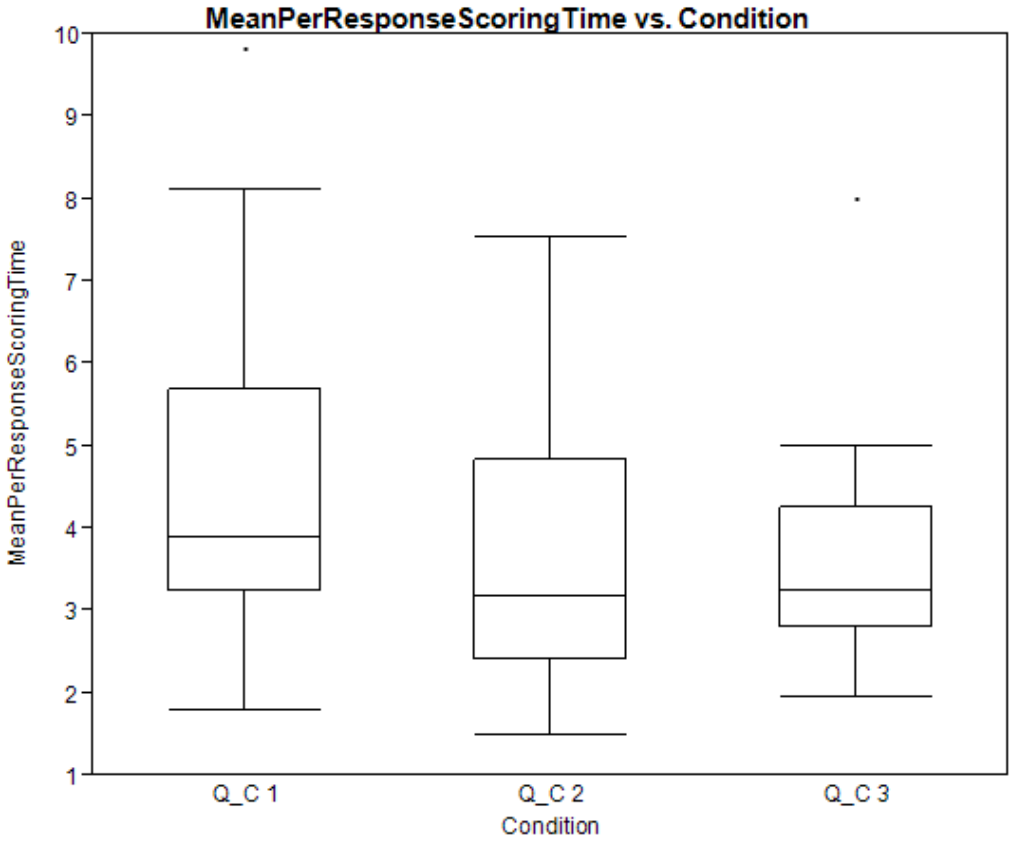
Scoring mode	Per response time (Grand mean)	
TT	3s	20
Q_C3	4s	16
Q_C2	4s	15
Q_C1	5s	12
Q_F	7s	7

responses per minute
(*in the given condition*)

TT marginally different from Q_C3

Time-on-task

Is scoring sheets of clustered responses faster than of non-clustered responses?



Time-on-task

Is total sheet scoring time linerly related to amount of material to score?

Time-on-task

Is total sheet scoring time linerly related to amount of material to score?

linear fit not statistical in the TT and Q_F conditions
large differences in RMSE in the Q conditions
best fit for Q_C2

Time-on-task

Is total sheet scoring time linerly related to amount of material to score?

linear fit not statistical in the TT and Q_F conditions
large differences in RMSE in the Q conditions
best fit for Q_C2

→ strongly dependent on **properties** of content

Conclusions

LREC-14: With basic clustering algorithm mid-80% accuracy can be achieved with 40% of responses scored (on our dataset)

Scoring clustered response sheets proceeds faster than scoring non-clustered response sheets and is comparable to the familiar scoring mode

Clustering is a promising direction for computer-assisted scoring

Problems with the presented study

Data: Items not comparable across conditions

Timing: Inaccurate estimate of response scoring time

Further work

Modelling
- Other clustering algorithms, other similarity measures
- Linguistically-informed features
- Pre-scoring

User-interface
- Score entry
- Response/Cluster presentation

Addressing the major problems of the present study
- Data: items not comparable across conditions
- Timing: inaccurate estimate of response scoring time
- → Controlled timing experiment

Thank you

(contact: magdalena.wolska@uni-tuebingen.de)