Document Understanding Conference
DUC 2006

Welcome!

DUC 2006-2007 Program Committee

John Conroy	IDA/CCS
Hoa Dang	NIST
Donna Harman	NIST
Ed Hovy	ISI/USC
Kathy McKeown	Columbia University
Drago Radev	University of Michigan
Karen Sparck-Jones	University of Cambridge
Lucy Vanderwende	Microsoft Research

DUC 2006 Agenda

```
================================================
Thursday, June 8
================================================
 9:00 -  9:15  Welcome/Intro
 9:15 - 10:00  Overview of task and NIST evaluation
10:00 - 10:30  Overview of Pyramid evaluation
10:30 - 11:00  B r e a k --------------------------
11:00 - 11:20  System talk: Simon Fraser University
11:20 - 11:40  System talk: Microsoft Research
11:40 - 12:00  System talk: LIA-Thales
12:00 - 12:20  Poster/boaster
12:30 -  2:00  L u n c h
 2:00 -  3:30  Group timeline exercise 1, discussion
 3:30 -  4:00  B r e a k --------------------------
 4:00 -  5:00  Group timeline exercise 2, discussion
 5:00 -  5:30  Plans for DUC 2007 and beyond
================================================
================================================
Friday, June 9
================================================
 9:00 -  9:20  System talk: IDA/CCS
 9:20 -  9:40  System talk: IIIT-Hyerabad
 9:40 - 10:00  System talk: Language Computer Corporation
10:00 - 10:20  System talk: Thomson Legal Research
10:30 - 11:00  B r e a k --------------------------
11:00 - 11:20  System talk: Columbia University
11:20 - 11:40  System talk: University of Twente
11:40 - 12:00  System talk: OGI-OHSU
12:10          Conclusion
```

Overview of DUC 2006

Evaluation of Question-Focused Text Summarization Systems

Hoa Dang

National Institute of Standards and Technology

June 8, 2006

Overview

- DUC background
- DUC 2006 framework
  - Task: documents, topics, model summaries
  - Manual evaluation: measures, procedures
- Results of DUC 2006 manual evaluation
  - Performance of peers based on various measures
  - Relation between measures
- Automatic evaluation of content
  - Correlation with manual evaluation
  - Comparison to DUC 2005
- Conclusion

Document Understanding Conferences (DUC)

- Originated out of TIDES program
- Summarization roadmap created in 2000, progress from:
  - simple genre → complex genre
  - simple tasks → demanding tasks
    - extract → abstract
    - single document → multiple documents
    - English → other language
    - generic summaries → focused or evolving summaries
  - intrinsic evaluation → extrinsic evaluation

DUC 2001-2005 investigated summarising:

- for single documents, multi-documents
- for news material
- at various lengths
- of various sorts including generic author-reflecting, viewpoint-oriented, novelty capturing, query-oriented
- comparing system summaries with manual ones, and (automatic) baseline ones
- using a range of evaluation criteria and performance measures including:
  - intrinsic measures: quality, coverage of reference summary content units (SEE; Pyramids), ngram coincidence with reference summary (ROUGE/BE)
  - extrinsic measures (simulated): usefulness and responsiveness.

DUC 2006 question-focused summarization task

- Given topic statement, document set
- Create fluent, 250-word answer to questions in topic statement, using information in document set

- Example topic statement:

  num: D0641E

  title: global warming

  narr: Describe theories concerning the causes and effects of global warming and arguments against these theories.

DUC 2006 topics, document sets, model summaries

- 50 topics developed by 9 NIST assessors
- Each topic consists of:
  - Topic statement: a set of questions or other expression of information need
  - Document set: 25 documents that contribute to answering the question(s) in the topic statement
- Documents from *Associated Press*, *New York Times*, and *Xinhua* newswire
- Model summaries written by 10 assessors (including 9 topic developers)
  - 4 model summaries per topic
  - About 4 hrs/summary

Example manual summary (D0641E)

As early as 1968 scientists suggested that global warming might cause disintegration of the West Antarctic Ice Sheet. Greenhouse gas emissions created by burning of coal, gas and oil were believed by most atmospheric scientists to cause warming of the Earth's surface which could result in increased frequency and intensity of storms, floods, heat waves, droughts, increase in malaria zones, rise in sea levels, northward movement of some species and extinction of others.

Some scientists, however, argued that there was no real evidence of global warming and others accepted it as a fact but attributed it to natural causes rather than human activity. In 1998 a petition signed by 17,000 U.S. scientists concluded that there is no basis for believing (1) that atmospheric CO2 is causing a dangerous climb in global temperatures, (2) that greater concentrations of CO2 would be harmful, or (3) that human activity leads to global warming in the first place.

By 1999 an intermediate position emerged attributing global warming to a shift in atmospheric circulation patterns that could be caused by either natural influences such as solar radiation or human activity such as CO2 emissions.

By 2000 opponents of programs to cut back greenhouse emissions admitted that there was evidence of global warming but questioned its cause and dire consequences. Proponents of plans to control emissions to a large extent admitted that the size of the human contribution to global warming is not yet known.

Participants and automatic runs in DUC 2006

ID	Organization	ID	Organization
1	(NIST baseline)	19	Universitat Politecnica de Catalunya
2	Oregon Health & Science University	20	University of Karlsruhe
3	Chinese Academy of Sciences	21	Fitchburg State College
4	CL Research	22	Hong Kong Polytechnic University
5	Columbia University	23	Peking University
6	Fudan University	24	International Institute of Information Technology
7	Information Sciences Institute (Zhou)	25	University College Dublin
8	IDA CCS *and* University of Maryland JIKD	26	Information Sciences Institute (Daume)
9	Macquarie University	27	Language Computer Corporation
10	Microsoft Research	28	University of Avignon
11	NK Trust, Inc.	29	Larim Unit (MIRACL Laboratory)
12	National University of Singapore	30	Tokyo Institute of Technology *and* Universidad Autonoma de Madrid
13	Simon Fraser University	31	Thomson Legal & Regulatory
14	Toyohashi University of Technology	32	University of Maryland *and* BBN Technologies
15	IDA Center for Computing Sciences	33	University of Michigan
16	University of Connecticut	34	University of Salerno
17	National Central University	35	University of Ottawa
18	University of Twente		

Baseline: First complete sentences (up to 250 words) of text field of most recent document

Evaluation methods

- Manual Evaluation:
  - Linguistic quality
  - Content
    - Content Responsiveness
    - Pyramids
  - Overall Responsiveness
- Automatic Evaluation of Content:
  - ROUGE/BE

Manual scoring scale

- 7 scores per summary (5 linguistic qualities, 1 content responsiveness, 1 overall responsiveness)
- Each score based on a 5-point scale
    1. Very poor
    2. Poor
    3. Barely acceptable
    4. Good
    5. Very good

Linguistic quality questions

Q1. Grammaticality: The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

Q2. Non-redundancy: There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., “Bill Clinton”) when a pronoun (“he”) would suffice.

Linguistic quality questions

Q3. Referential clarity: It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

Q4. Focus: The summary should have a focus; sentences should only contain information that is related to the rest of the summary.

Q5. Structure and Coherence: The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

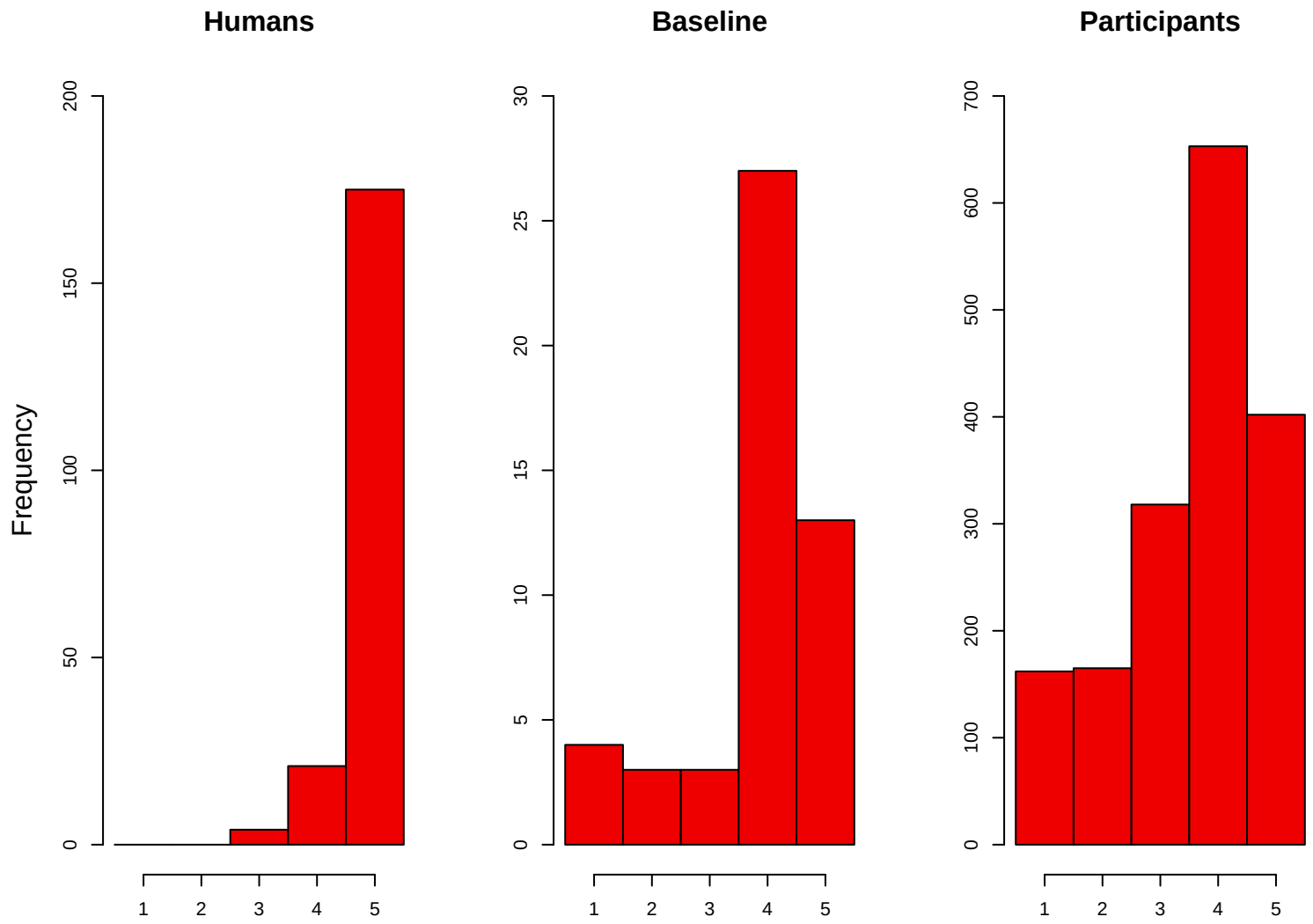
Responsiveness

- Content responsiveness
  - based on amount of information in summary that contributes to meeting the information need expressed in the topic
  - different strategies for scoring content
- Overall responsiveness
  - based on both information content and readability
  - “gut reaction” to summary
  - “How much would I pay for this summary?”

Manual assessment

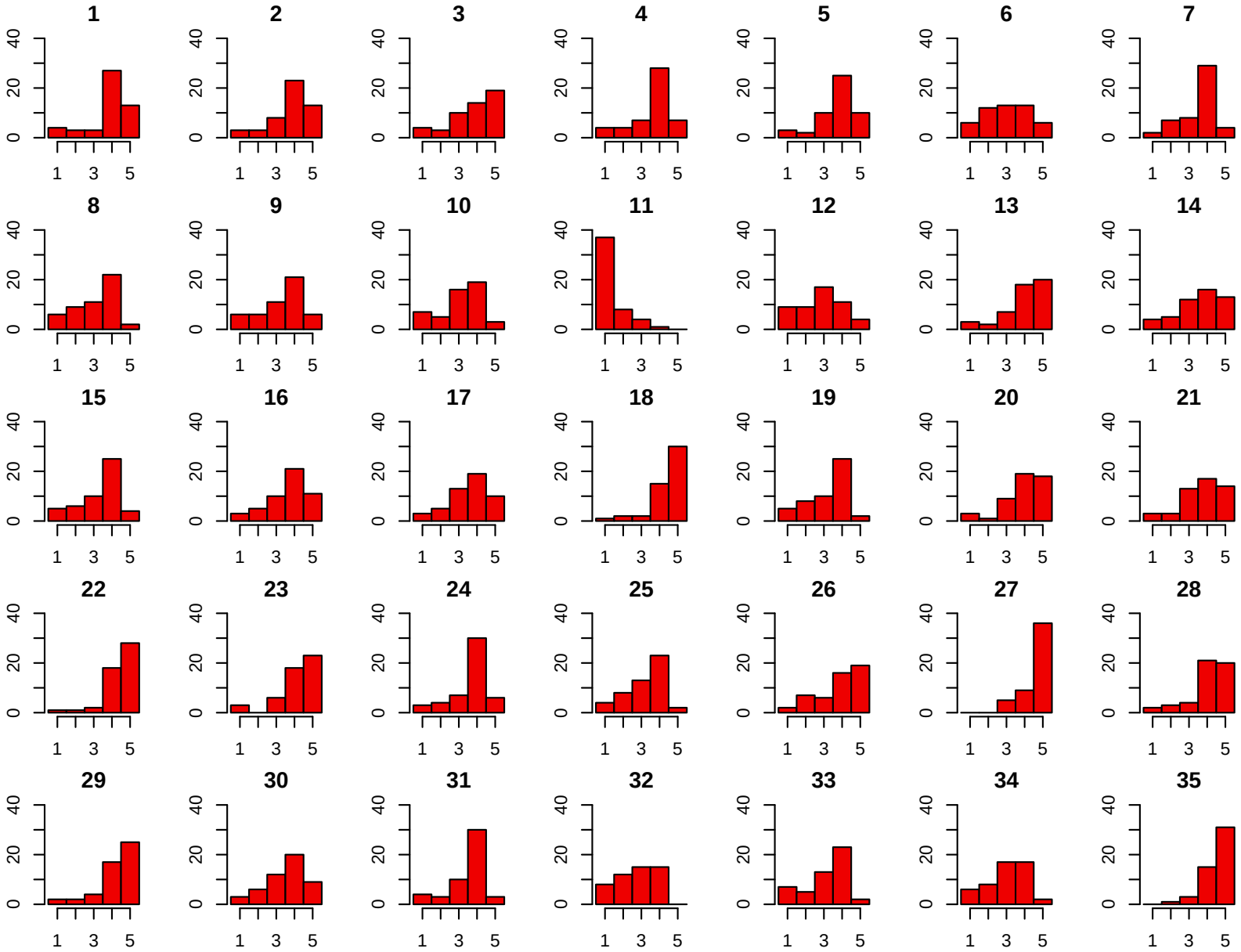
- 10 Assessors
- One assessor per topic: Linguistic quality, content responsiveness, overall responsiveness
  - Assessor usually the same as topic developer
  - Assessor always one of the summarizers for the topic
- ```
  for each topic
    assess summaries for linguistic qualities
    assess summaries for content responsiveness
  foreach topic
    assess summaries for overall responsiveness
  ```
- 5 hours per topic (average)
  ```

Q1: Grammaticality

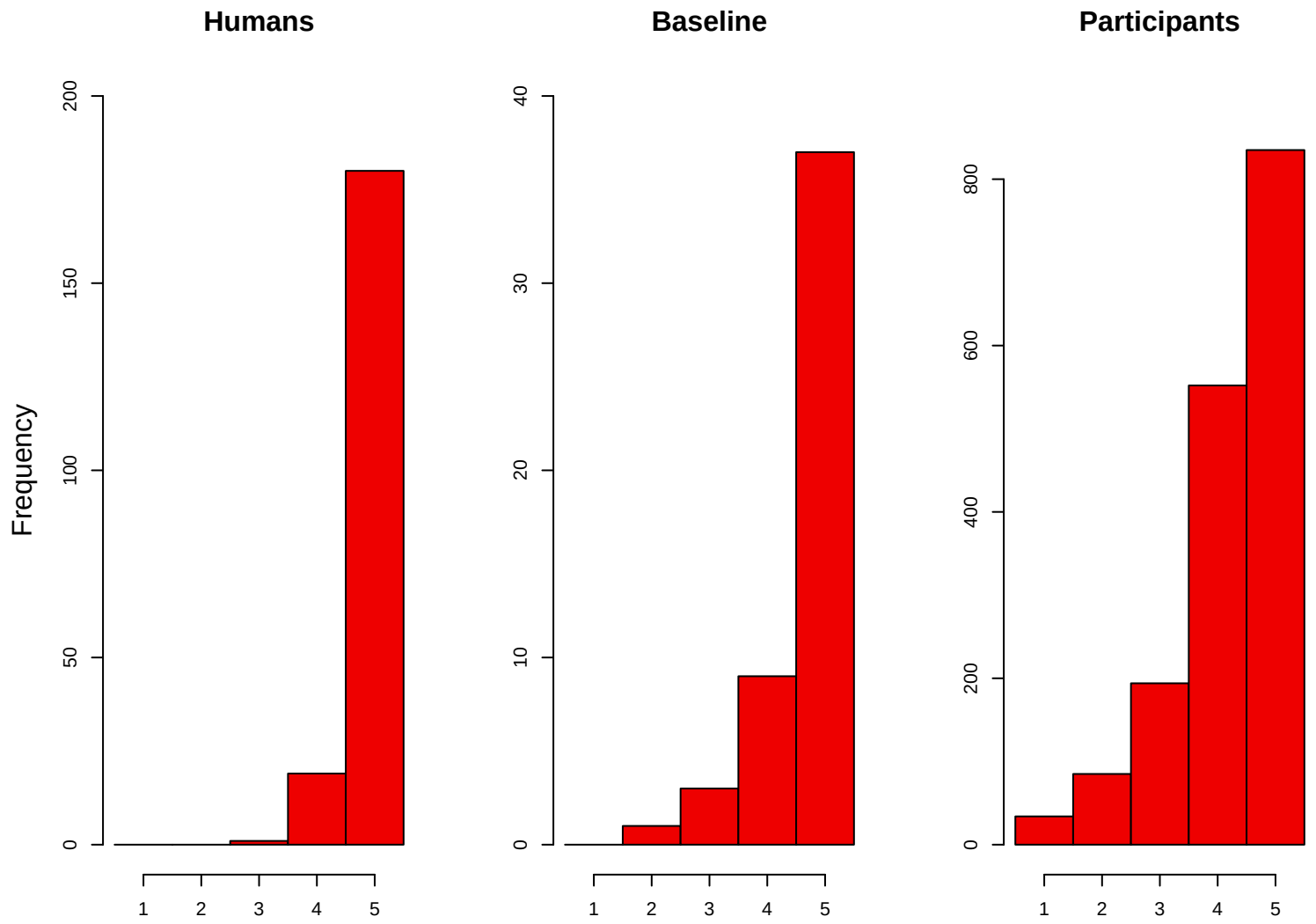


Similar to 2005

Q1: Grammaticality

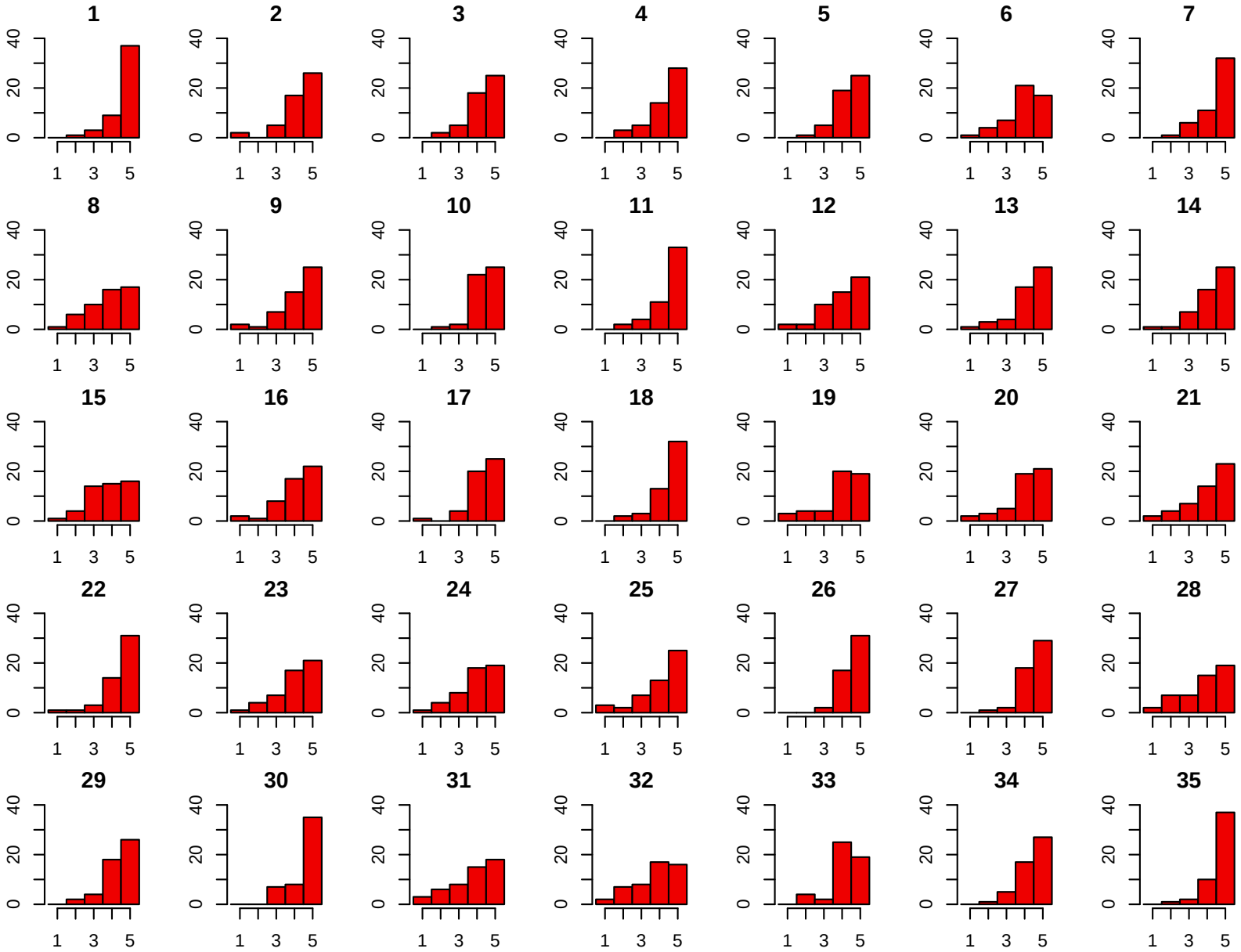


Q2: Non-redundancy

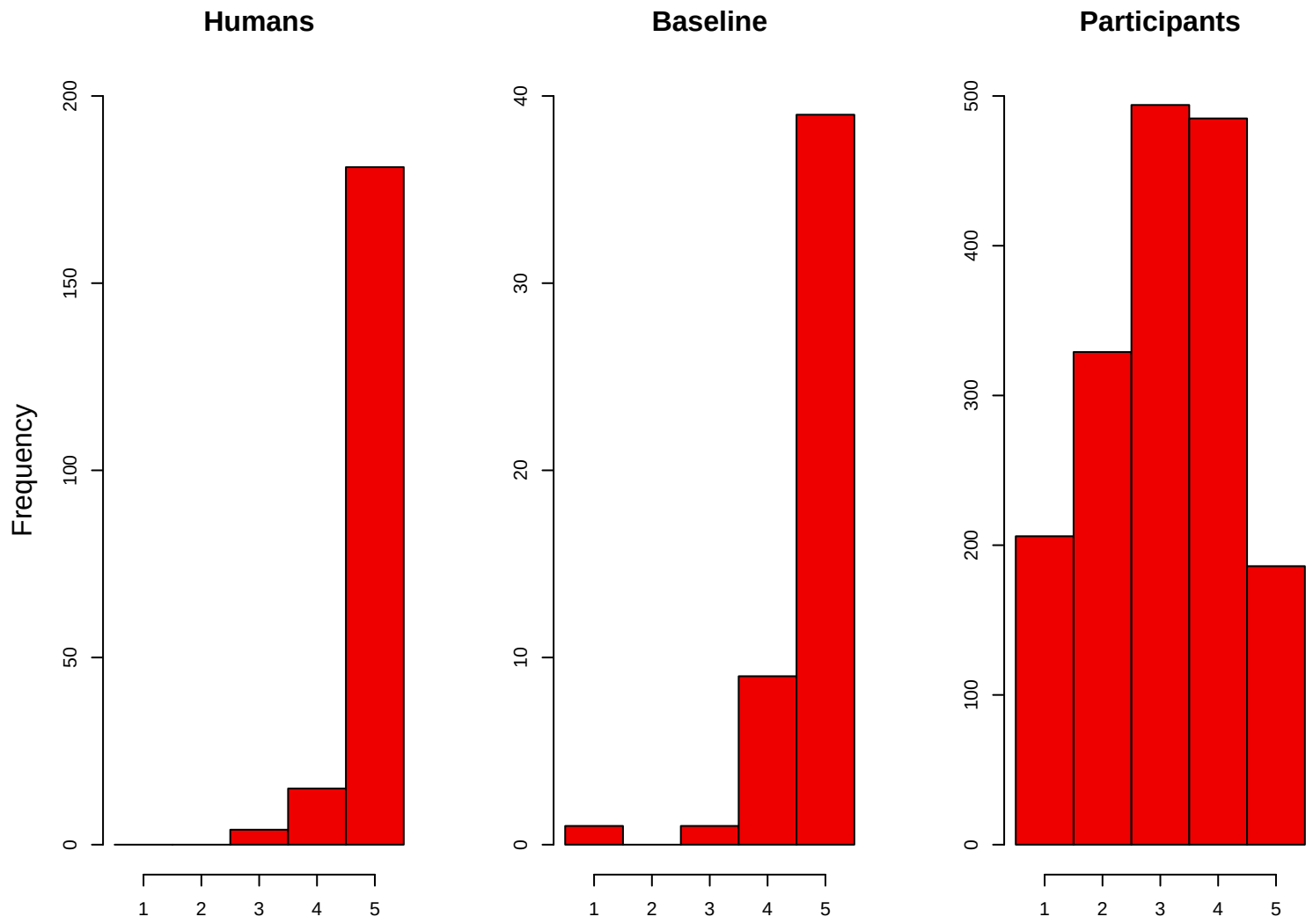


Similar to 2005

Q2: Non-redundancy

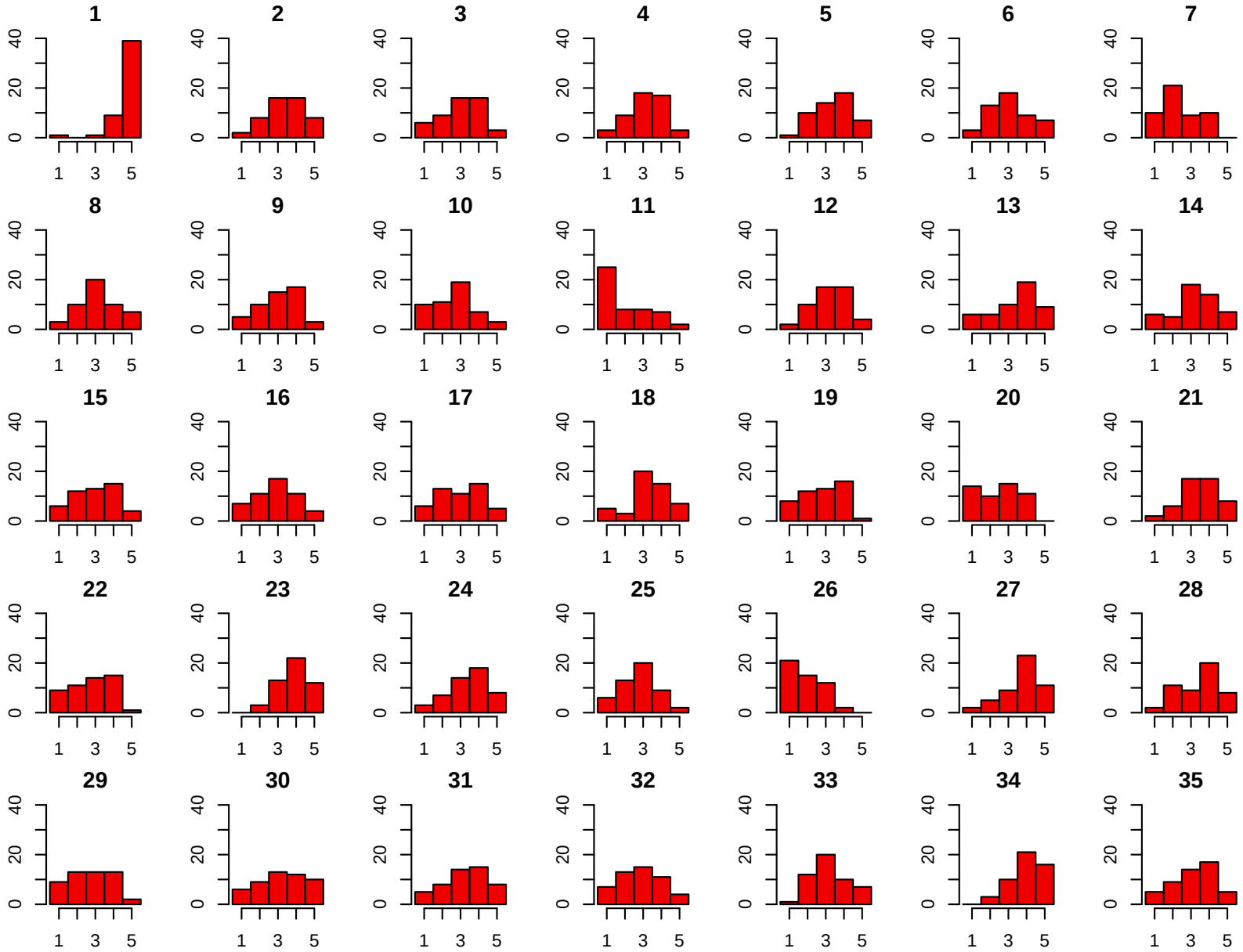


Q3: Referential clarity

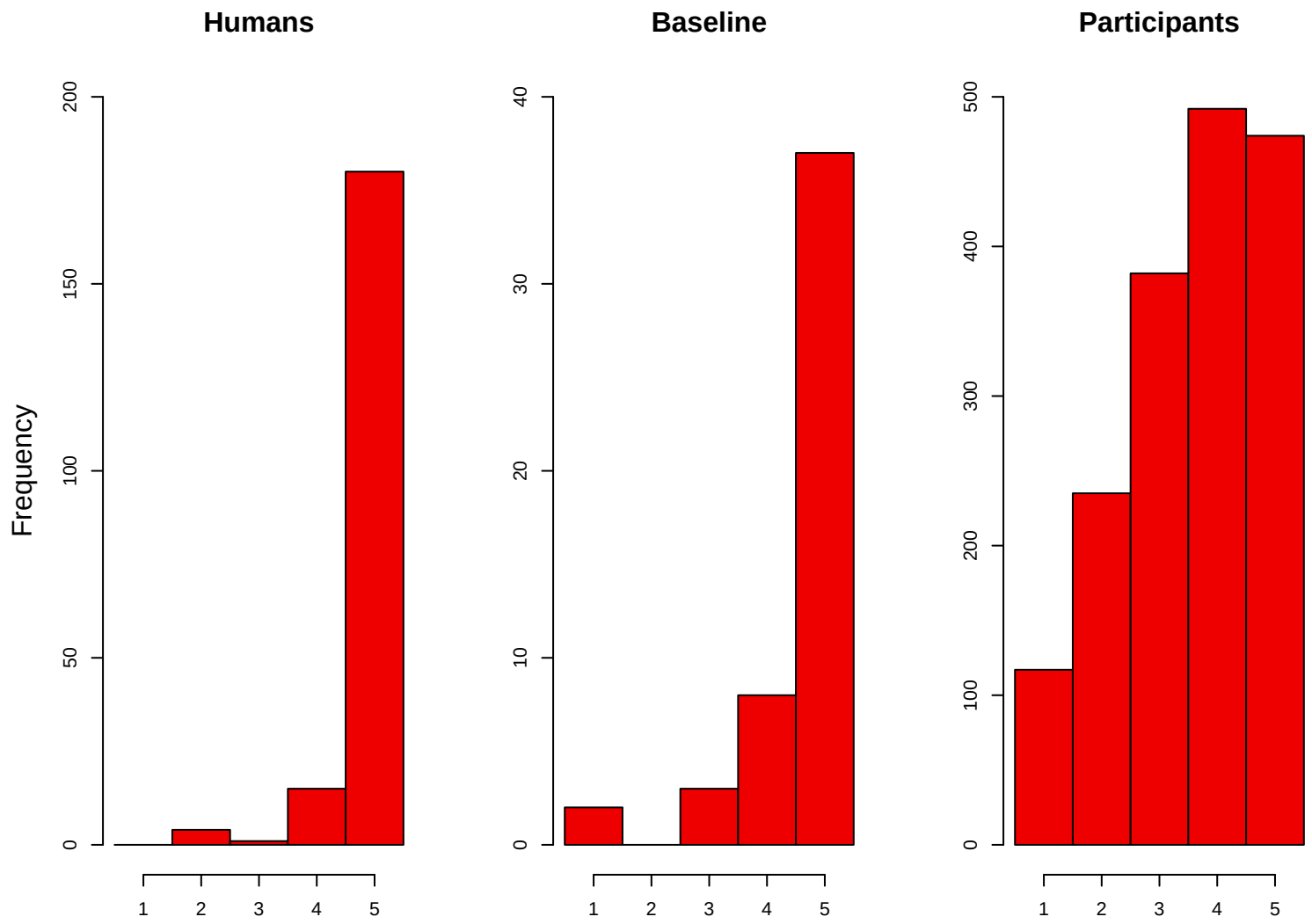


Similar to 2005

Q3: Referential clarity

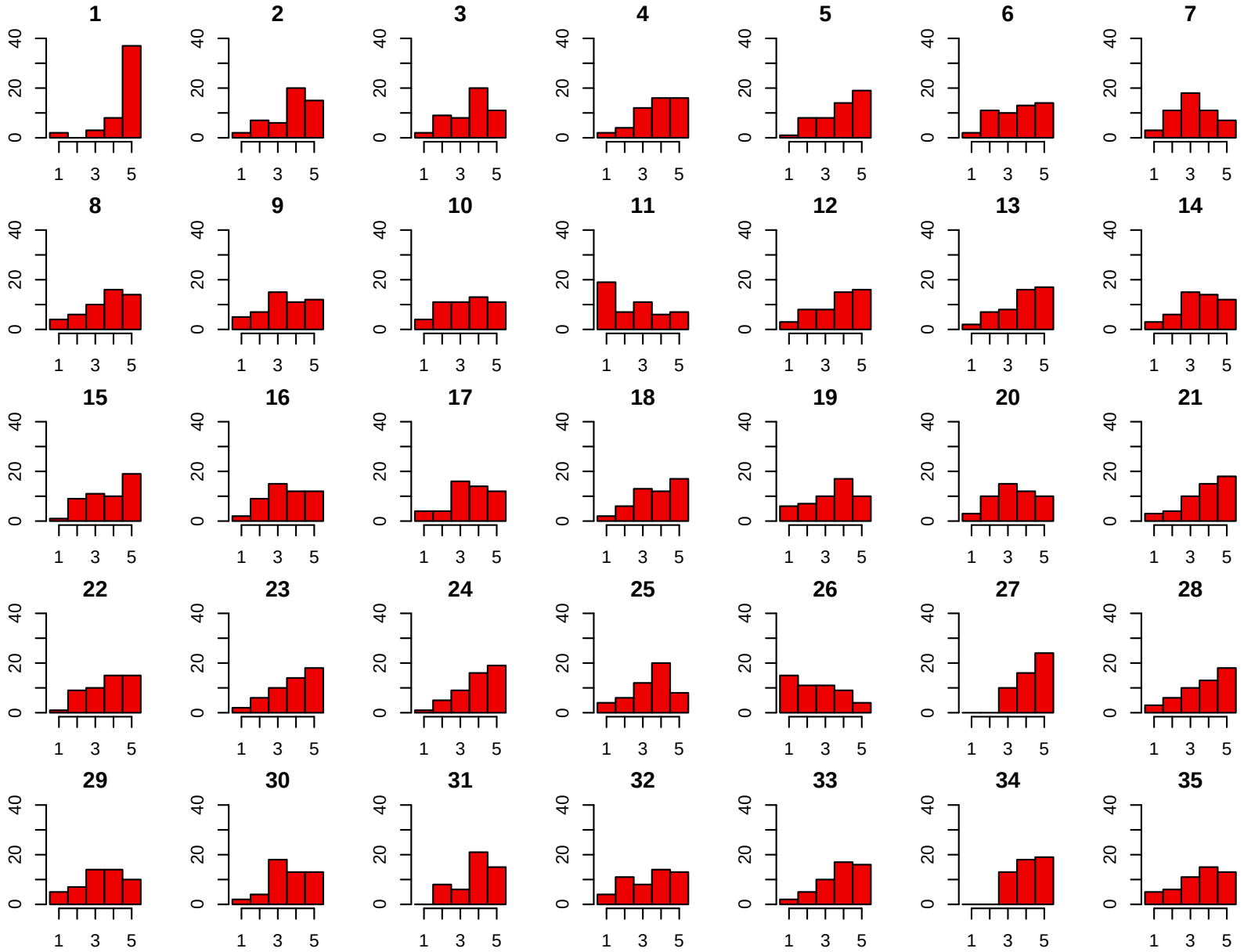


Q4: Focus



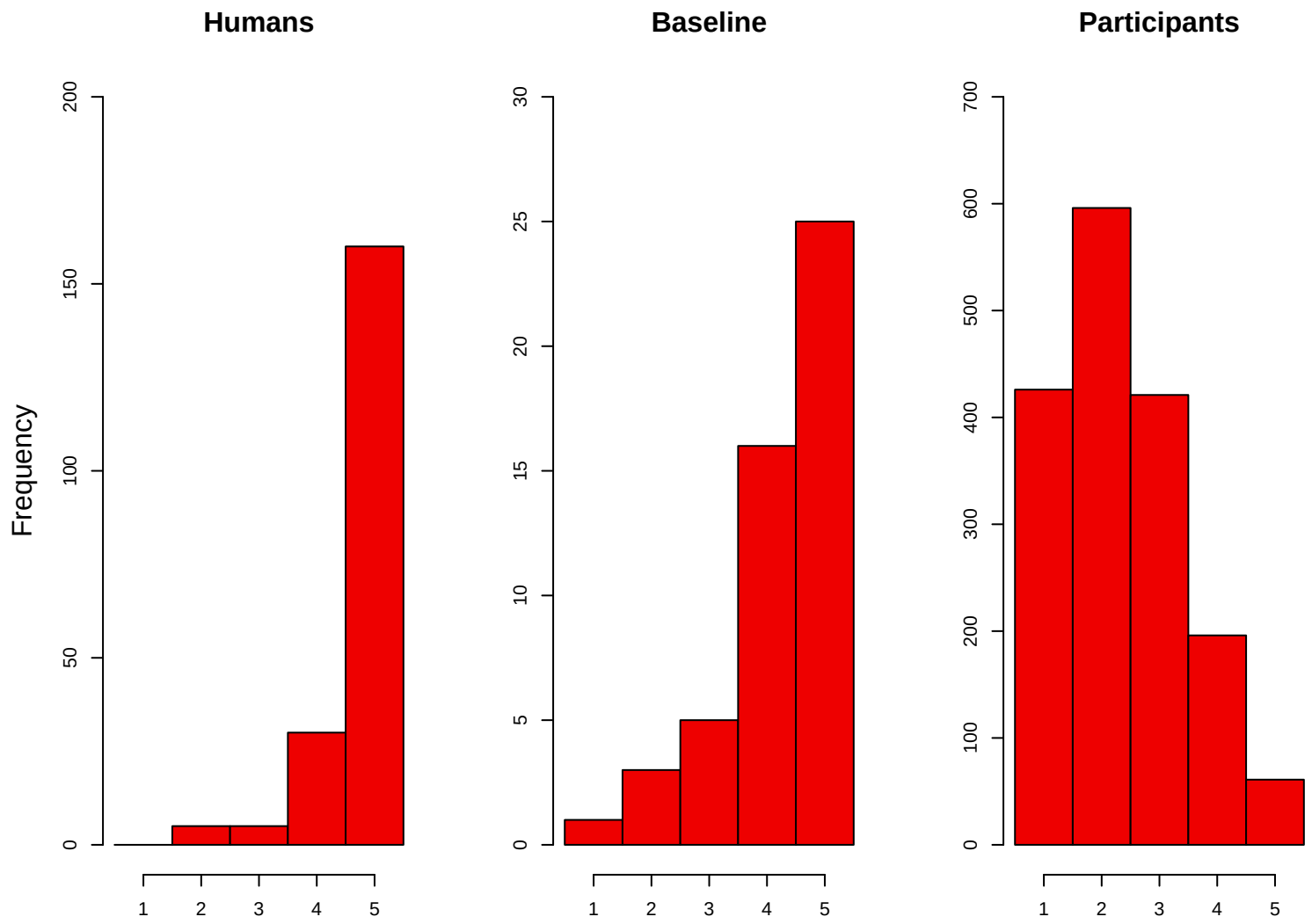
Better than 2005

Q4: Focus



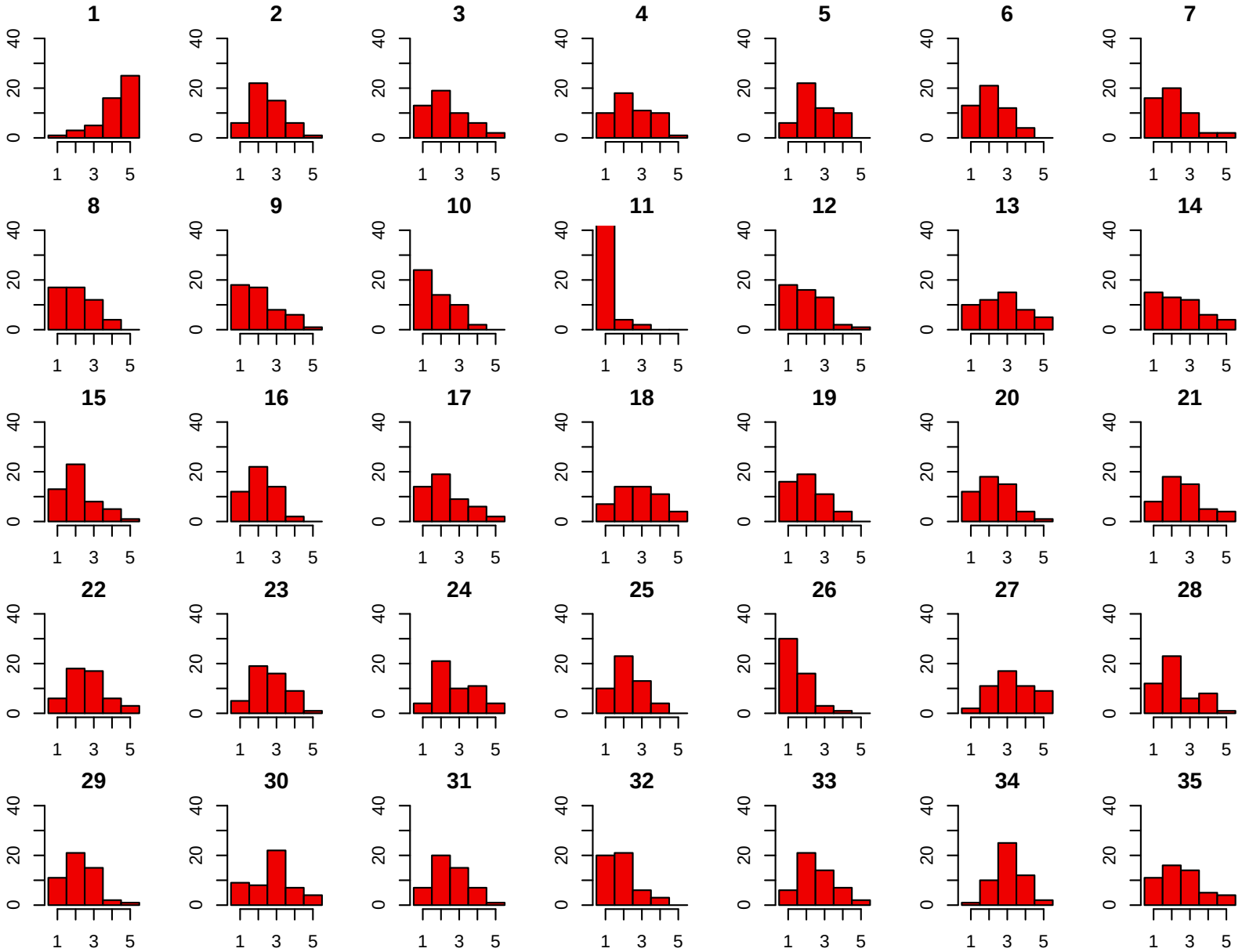
mostly good

Q5: Structure and coherence

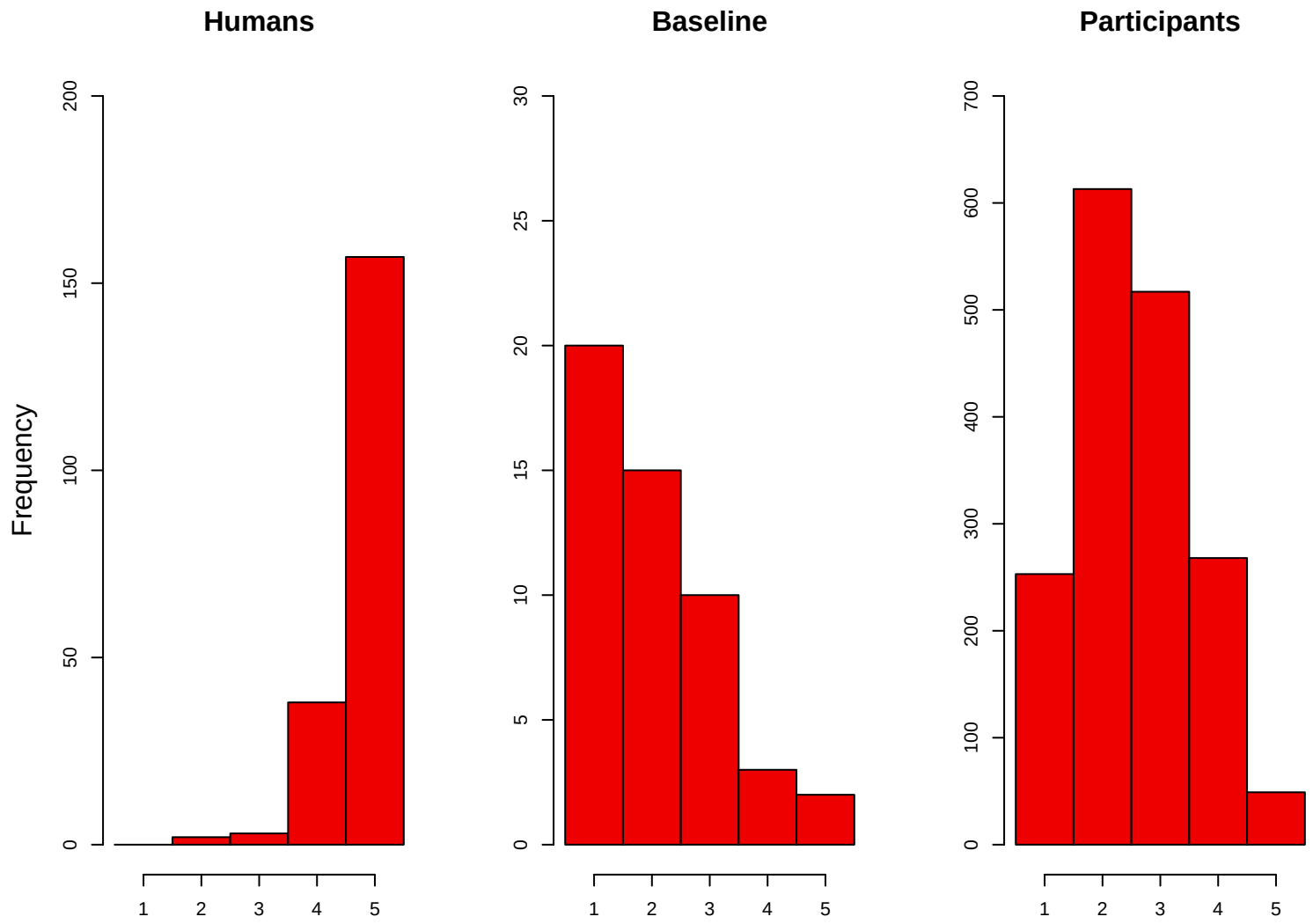


Similar to 2005

Q5: Structure and coherence

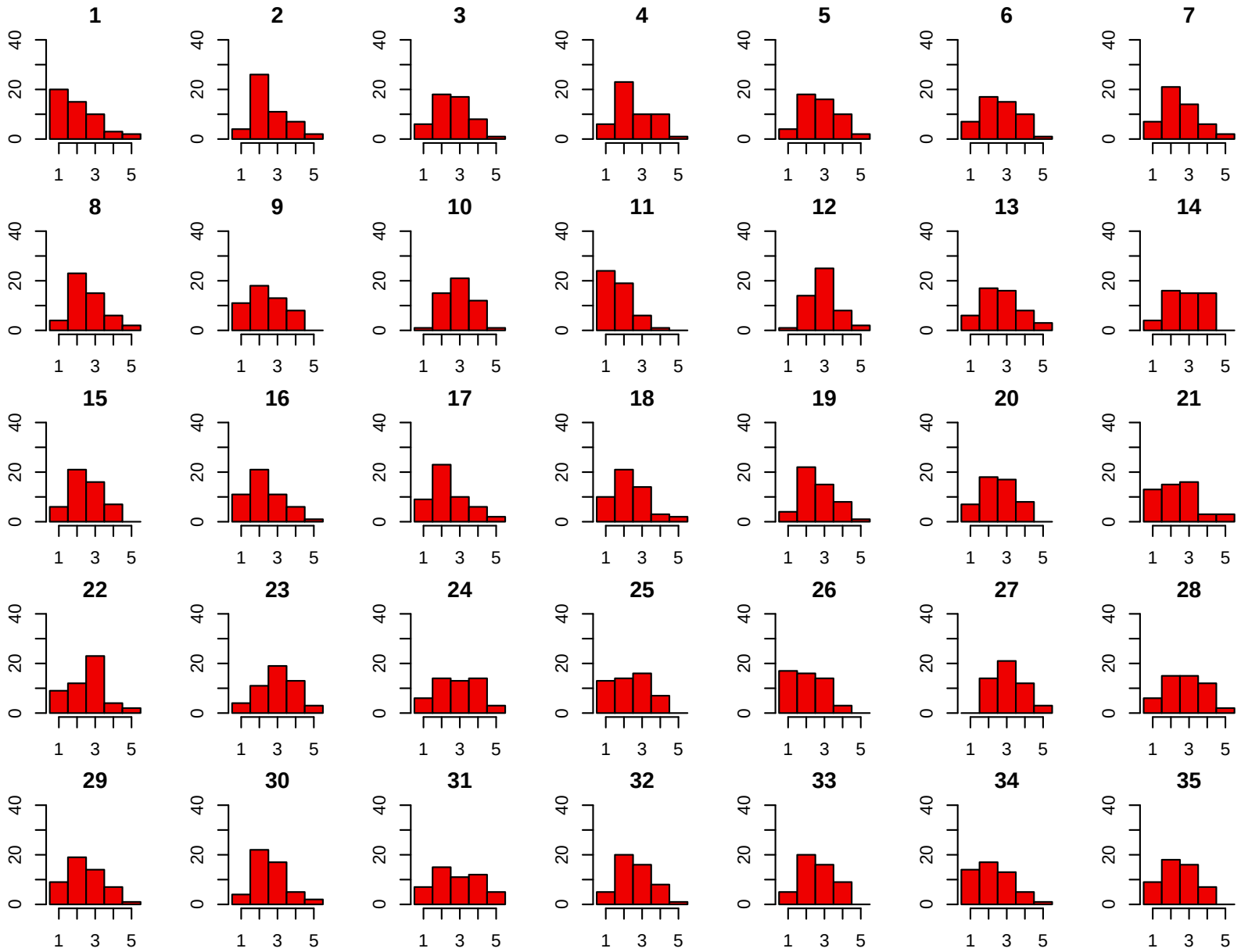


Content Responsiveness

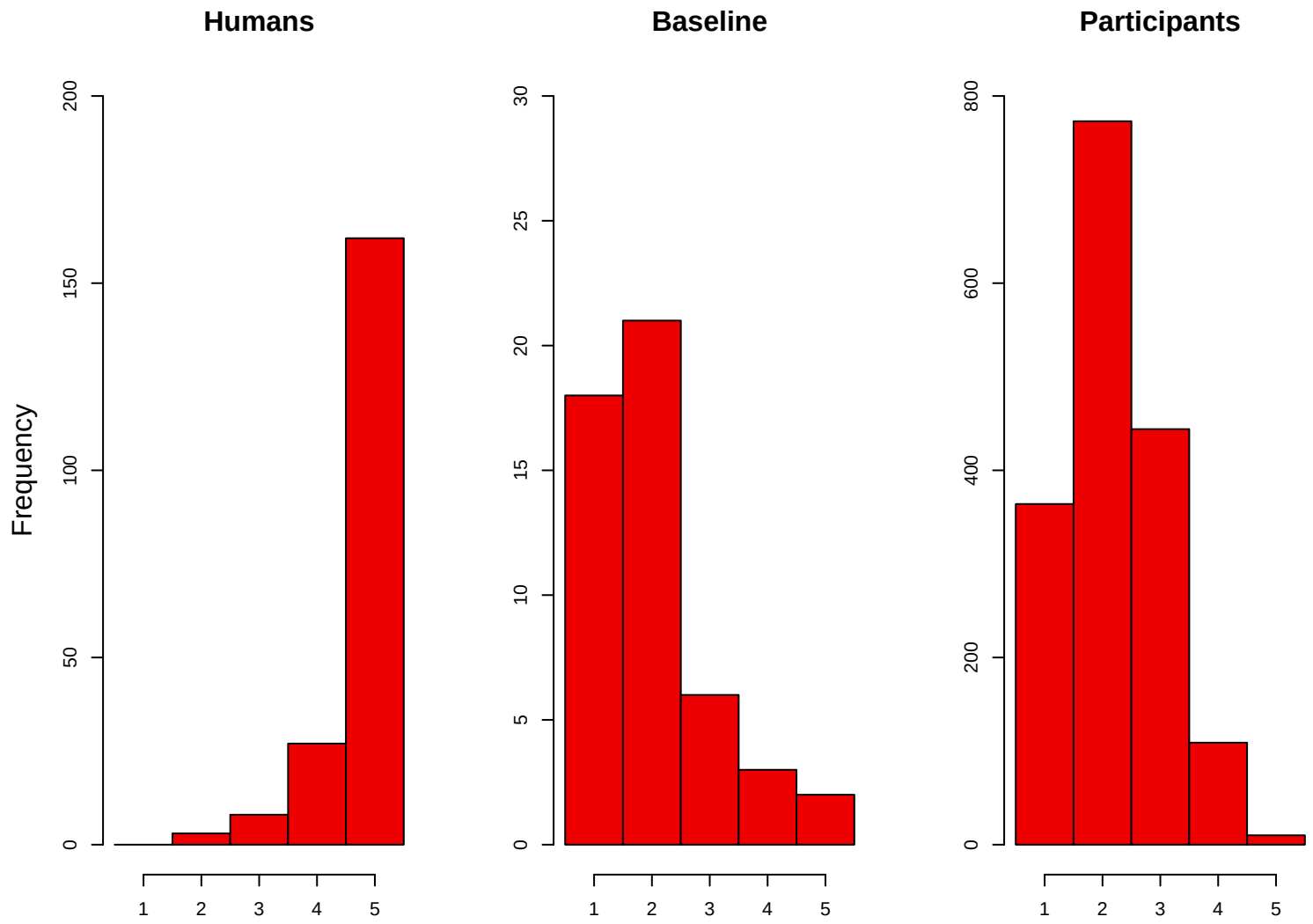


Much better than baseline

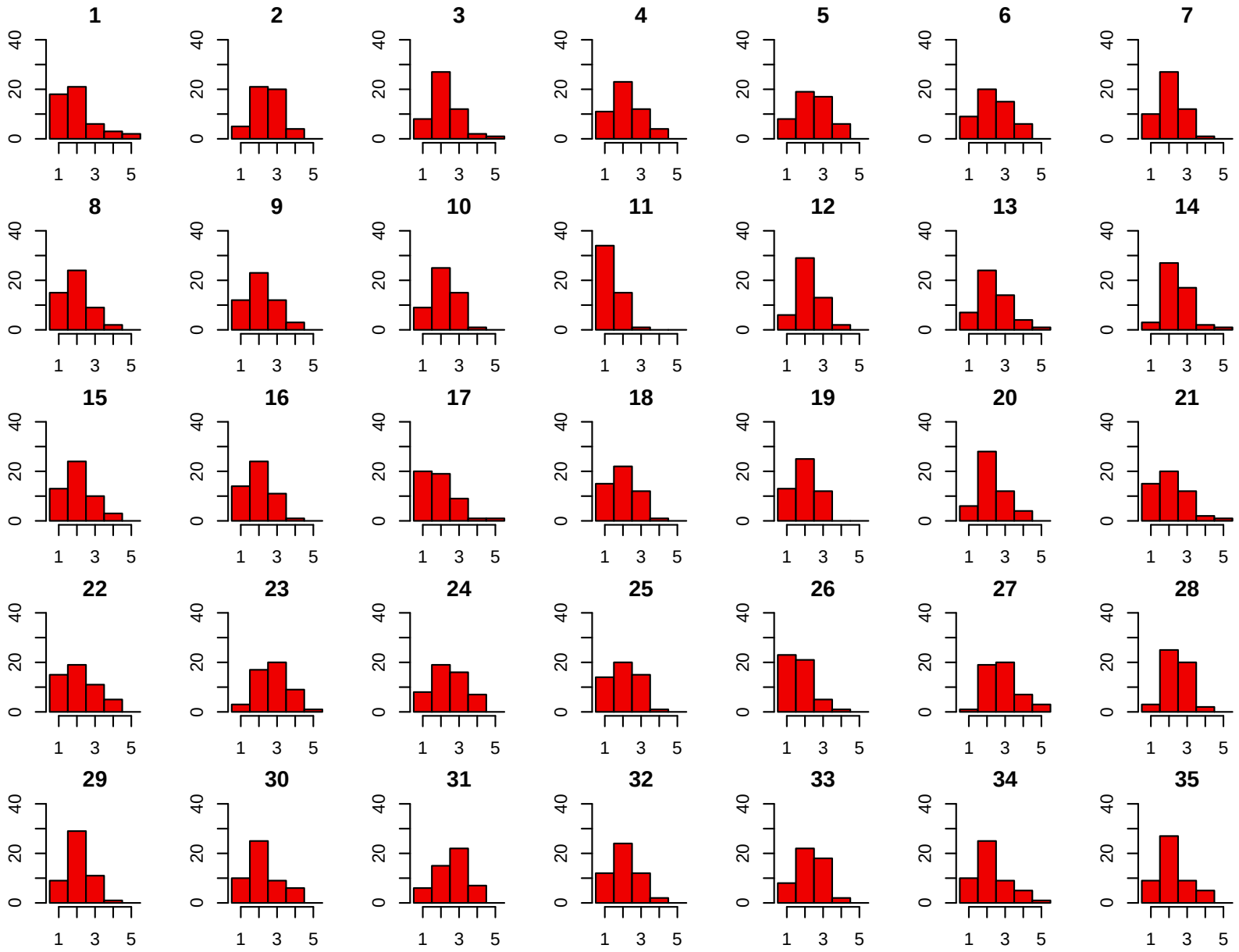
Content Responsiveness



Overall Responsiveness



Overall Responsiveness



Multiple Linear Regression: $y = \beta X + \epsilon$

The purpose of multiple linear regression is to establish a quantitative relationship between a group of predictor variables X and a response, y. This relationship is useful for:

- Understanding which predictors have the greatest effect.
- Knowing the direction of the effect (i.e., increasing x $\in$ X increases or decreases y).
- Using the model to predict future values of the response when only the predictors are currently known.

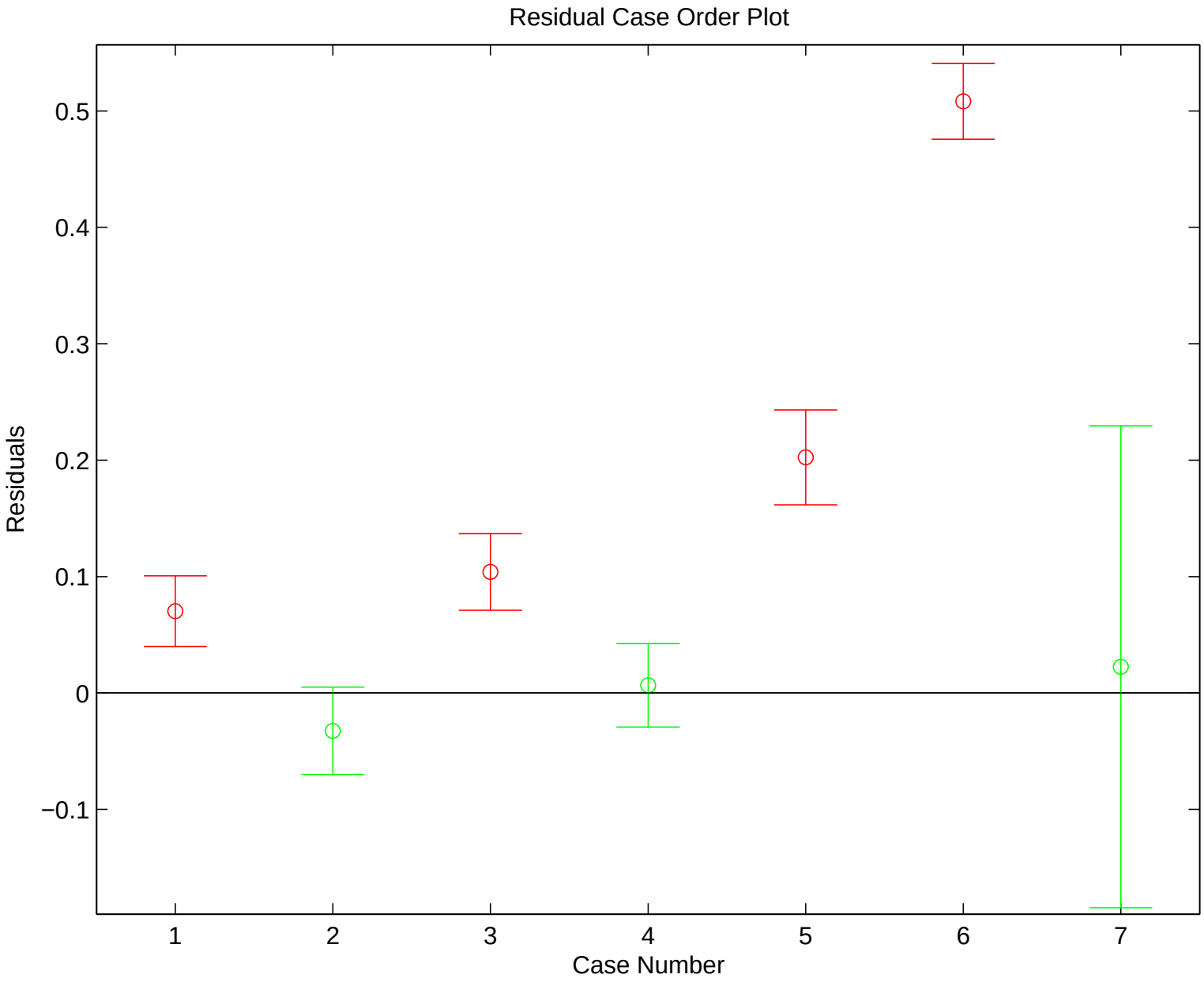
Examine effect of 5 linguistic qualities and content responsiveness on overall responsiveness

Multiple Regression

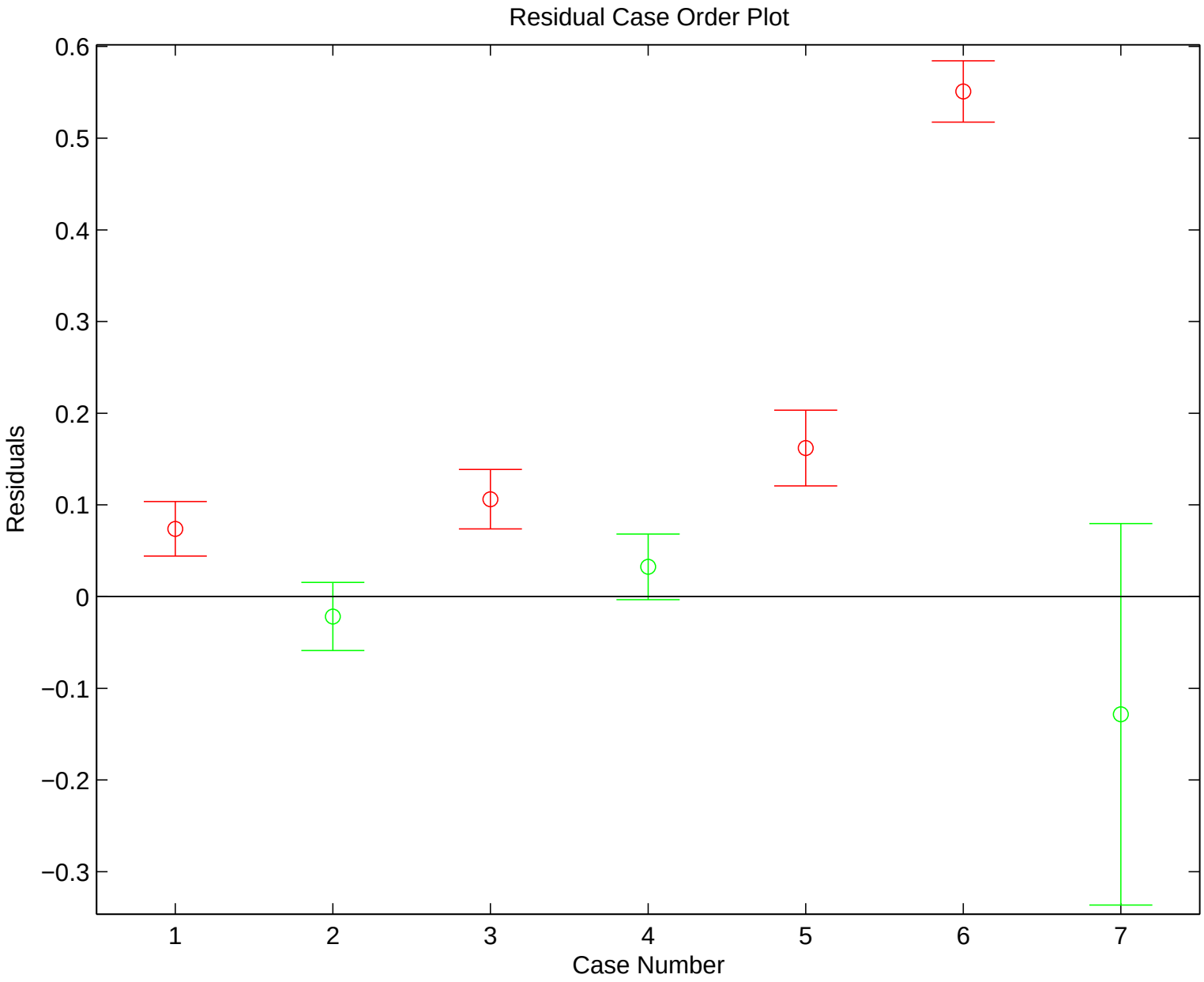
Assessor	Q1: β	Q2: β	Q3: β	Q4: β	Q5: β	content: β	R^2
B	0.0623	-0.1068	0.0604	-0.0738	0.2996	0.5955	0.7543
J	0.0419	0.0106	0.0355	-0.0902	0.4183	0.5366	0.7439
A	-0.0016	-0.0374	0.0560	0.0618	0.1033	0.6973	0.7316
E	0.0677	0.0153	0.2513	0.0463	0.0803	0.5028	0.6911
I	0.0789	0.0165	0.0969	-0.0135	0.1736	0.5765	0.6221
D	0.0207	-0.0289	0.0073	0.0129	0.3415	0.4936	0.5512
C	-0.0003	-0.0822	0.1695	-0.0223	0.1977	0.5474	0.5096
F	-0.0277	0.2280	0.1635	-0.0302	-0.0510	0.7250	0.4759
H	0.1018	-0.0169	0.1395	-0.1494	0.2569	0.3909	0.4530
G	0.0389	-0.1576	0.2211	-0.0286	0.5293	0.1967	0.3945

R^2 measures amount of the variability in the observations accounted for by the model

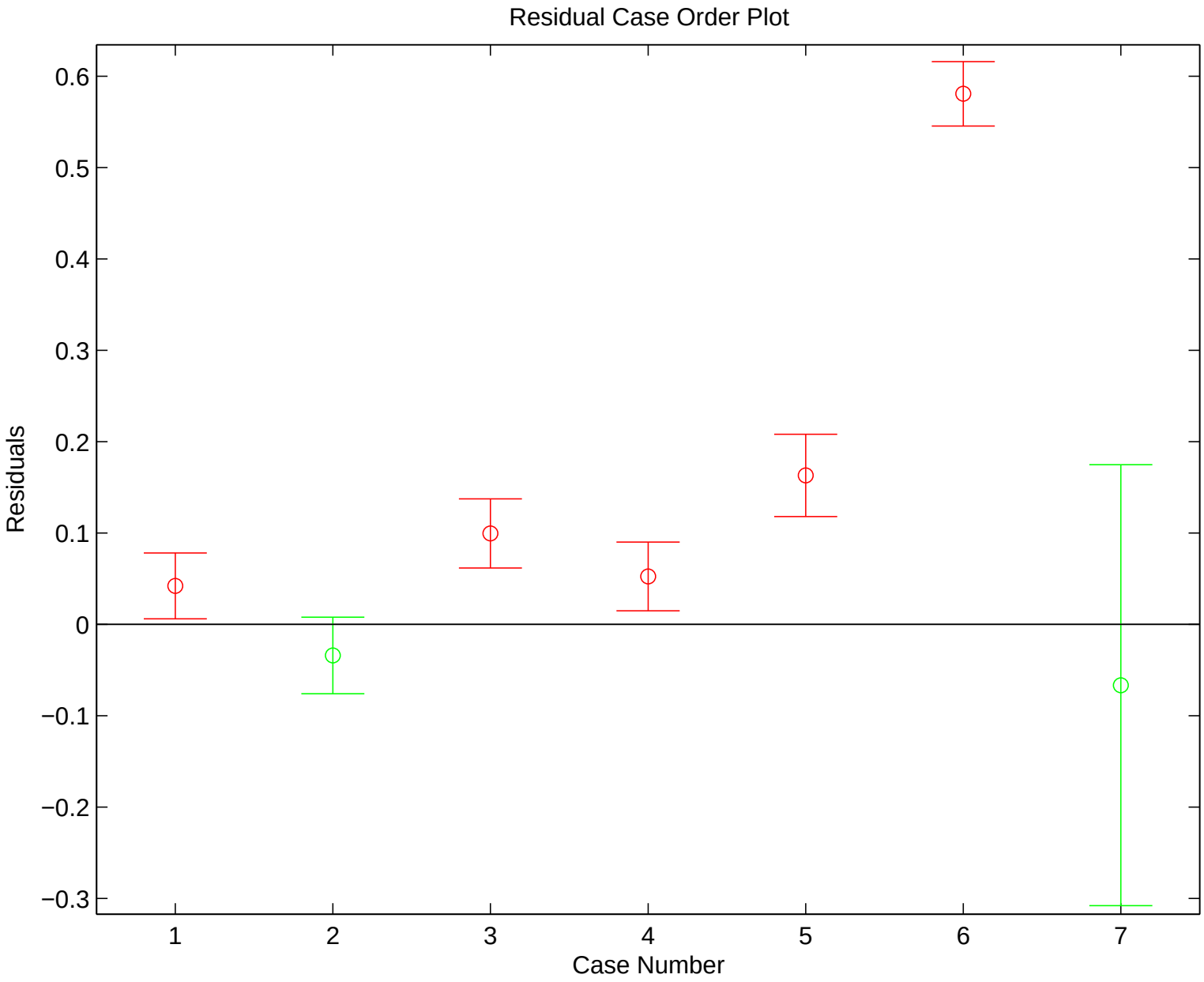
Multiple Regression, all 10 assessors: $R^2 = 0.5877$



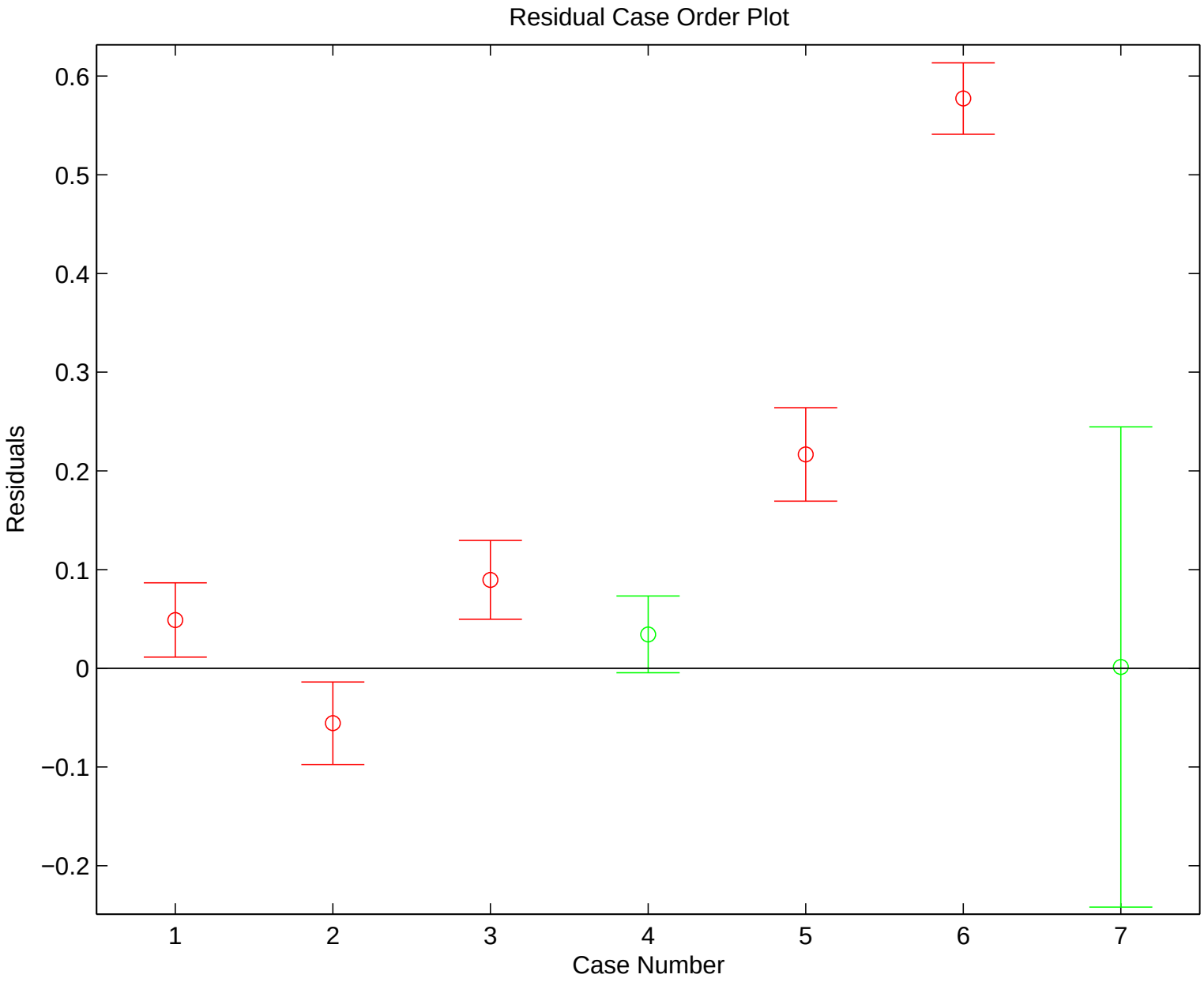
Multiple Regression, 9 assessors: $R^2 = 0.6031$



Multiple Regression, 8 assessors: $R^2 = 0.6284$



Multiple Regression, 7 assessors: $R^2 = 0.6647$



Past manual intrinsic evaluation of content (SEE Coverage)

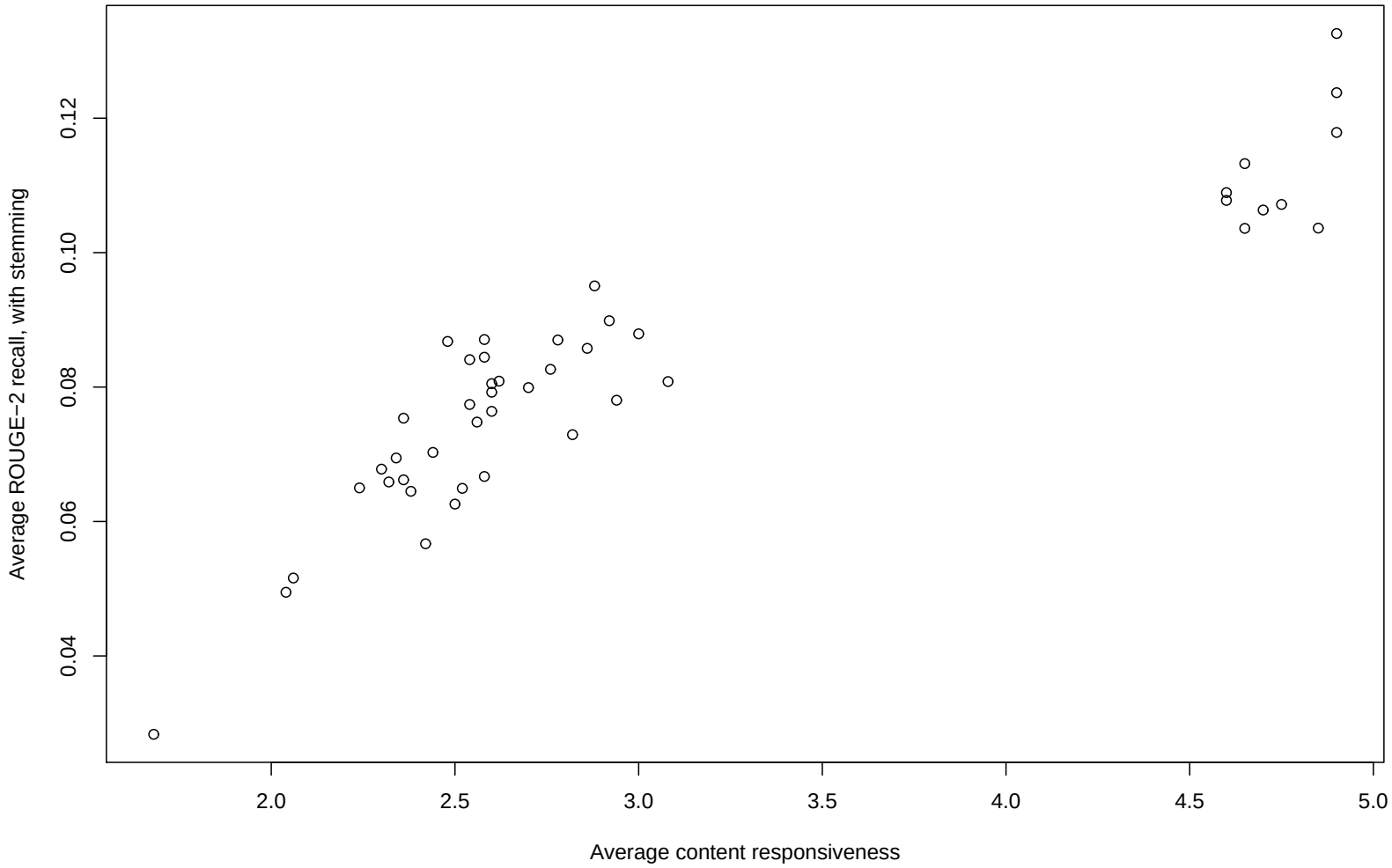
- Compare each peer (human or automatic) against a single model (human) summary
- Segment model summary into model units (MUs=EDU) (Soricut and Marcu, 2003)
- For each Model Unit:
  1. mark peer sentences that express any of the meaning in the MU
  2. the marked peer sentences together express [0, 20, 40, 60, 80, 100]% of the meaning expressed in the Model Unit.
- Mean coverage: average of the per-Model Unit judgments

Variation in human summaries $\rightarrow$
Compare against multiple model summaries (Pyramids/ROUGE)

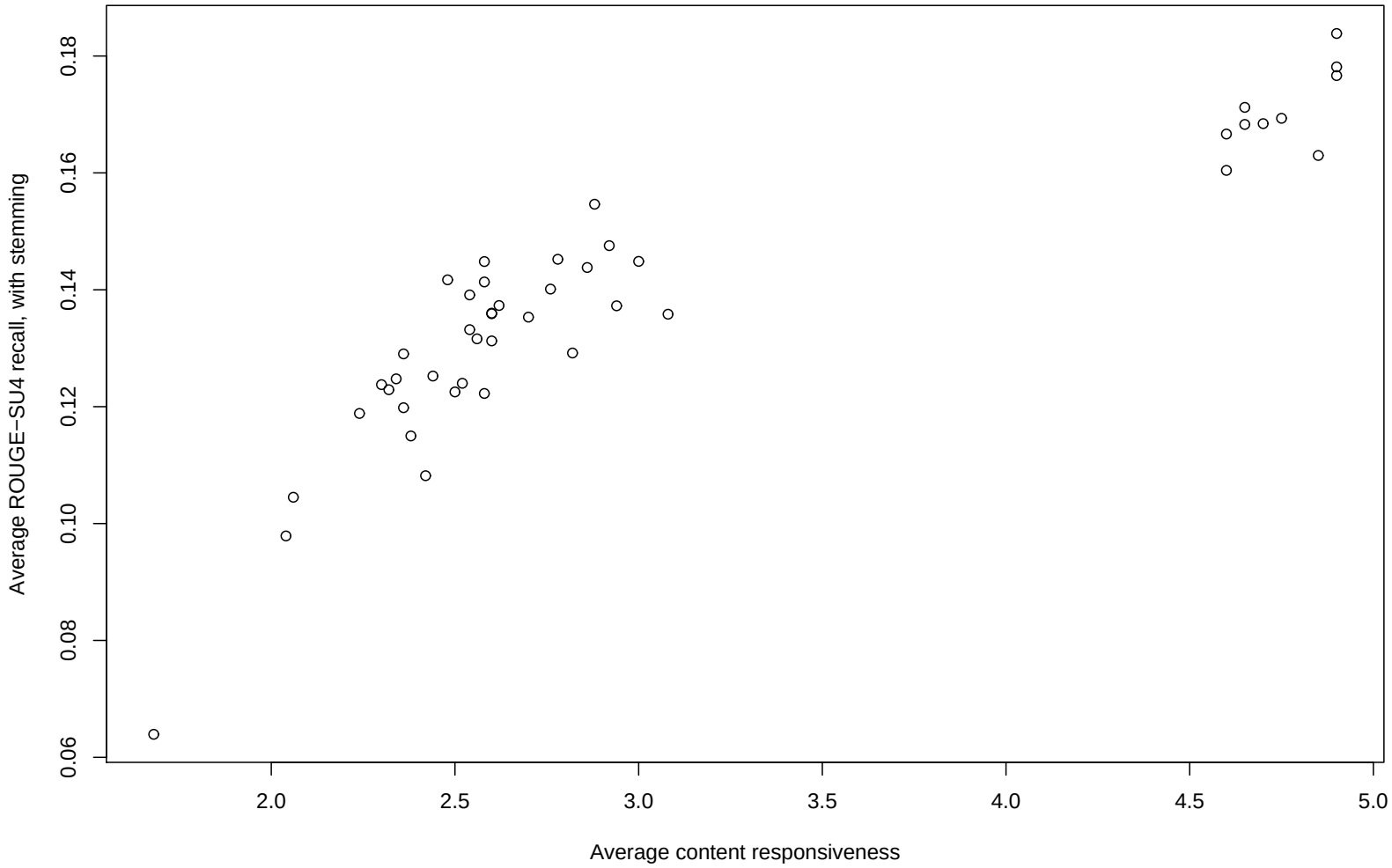
ROUGE-1.5.5

- Match n-grams between each peer summary and set of model summaries
- Weight n-gram by number of model summaries that it appears in
- Recall-oriented (precision, f-measure also available)
- DUC 2006 automatic metrics:
  - ROUGE-2: match word bigrams
  - ROUGE-SU4: match skip bigrams, with skip distance of up to 4 words: "the dog" = "the quick scary little brown dog"
  - Basic Elements: match head-modifier pairs extracted from automatic parse of summaries
- stem words before matching
- implement jackknifing for each (peer, topic) pair

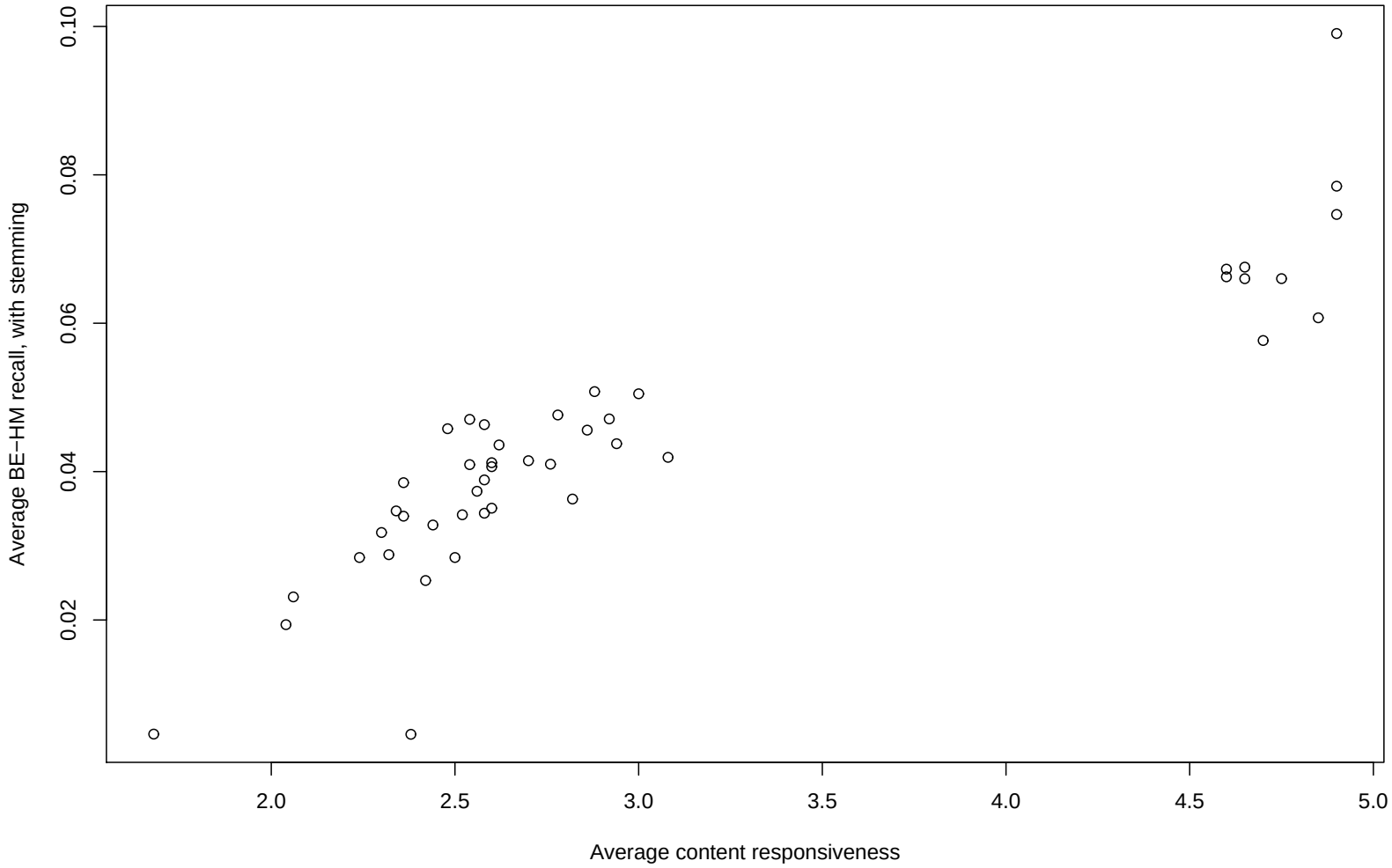
ROUGE-2 recall vs. Content Responsiveness



ROUGE-SU4 recall vs. Content Responsiveness



BE-HM recall vs. Content Responsiveness



Correlation with average content responsiveness

Metric	Spearman	Pearson
overall responsiveness	0.718	0.833 [0.720, 1.000]
ROUGE-2	0.767	0.836 [0.725, 1.000]
ROUGE-SU4	0.790	0.850 [0.746, 1.000]
BE-HM	0.797	0.782 [0.641, 1.000]

Correlations between manual and automatic content scores lower than in 2005.

Comparison with DUC 2005

	2005	2006
Corpus	LA/Financial Times	NYT, AP, Xinhua
Topics	old TREC topics	new
Docset size	25-50 (avg. 32)	25
Specific Granularity?	Yes	No
Models	4 or 9	4
Content Responsiveness	5 relative clusters	“absolute” scale

Correlations with average content responsivenss: 2005 and 2006

2006 Metric	Spearman	Pearson
ROUGE-2 (all topics)	0.759	0.835 [0.722, 1.000]
ROUGE-SU4 (all topics)	0.780	0.849 [0.745, 1.000]

2005 Metric	Spearman	Pearson
ROUGE-2 (all topics)	0.889	0.926 [0.868, 1.000]
ROUGE-SU4 (all topics)	0.867	0.917 [0.852, 1.000]

Correlations with average content responsivenss: 2005 and 2006

2006 Metric	Spearman	Pearson
ROUGE-2 (all topics)	0.759	0.835 [0.722, 1.000]
ROUGE-SU4 (all topics)	0.780	0.849 [0.745, 1.000]

2005 Metric	Spearman	Pearson
ROUGE-2 (all topics)	0.889	0.926 [0.868, 1.000]
ROUGE-SU4 (all topics)	0.867	0.917 [0.852, 1.000]
ROUGE-2 (general)	0.804	0.827 [0.702, 1.000]
ROUGE-SU4 (general)	0.841	0.868 [0.770, 1.000]
ROUGE-2 (specific)	0.912	0.928 [0.871, 1.000]
ROUGE-SU4 (specific)	0.884	0.921 [0.858, 1.000]

Conclusion

- Use caution if optimizing on automatic scores (esp., for general/abstractive summaries)
- Combination of content and readability is important for overall responsiveness
  - Focus and Non-Redundancy have less impact
- Automatic summaries are much improved...Good Job!