HOW REPRODUCIBLE IS BEST-WORST SCALING FOR HUMAN EVALUATION?

A REPRODUCTION OF

DATA-TO-TEXT WITH MACRO-PLANNING

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen,
Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, Emiel Krahmer

GOAL: REPRODUCE THIS EXPERIMENT AS CLOSELY AS POSSIBLE

Summaries

System Summaries
A: The Golden State Warriors (43 - 7) defeated the Los Angeles Clippers (31 - 19) 133 - 120 on Saturday. The Warriors came into this game as one of the best defenses in the NBA this season, but they were able to prevail with a huge road win. [. . . 11 more sentences]

B: The Golden State Warriors defeated the Los Angeles Clippers, 133 - 120, at Staples Center on Wednesday. The Warriors (43 - 7) came into this game as a sizable favorite and they showed why in this clincher. Golden State (31 - 19) came into this game as a huge favorite and they showed some resiliency here with this win. [. . . 11 more sentences]"

Ranking Criteria

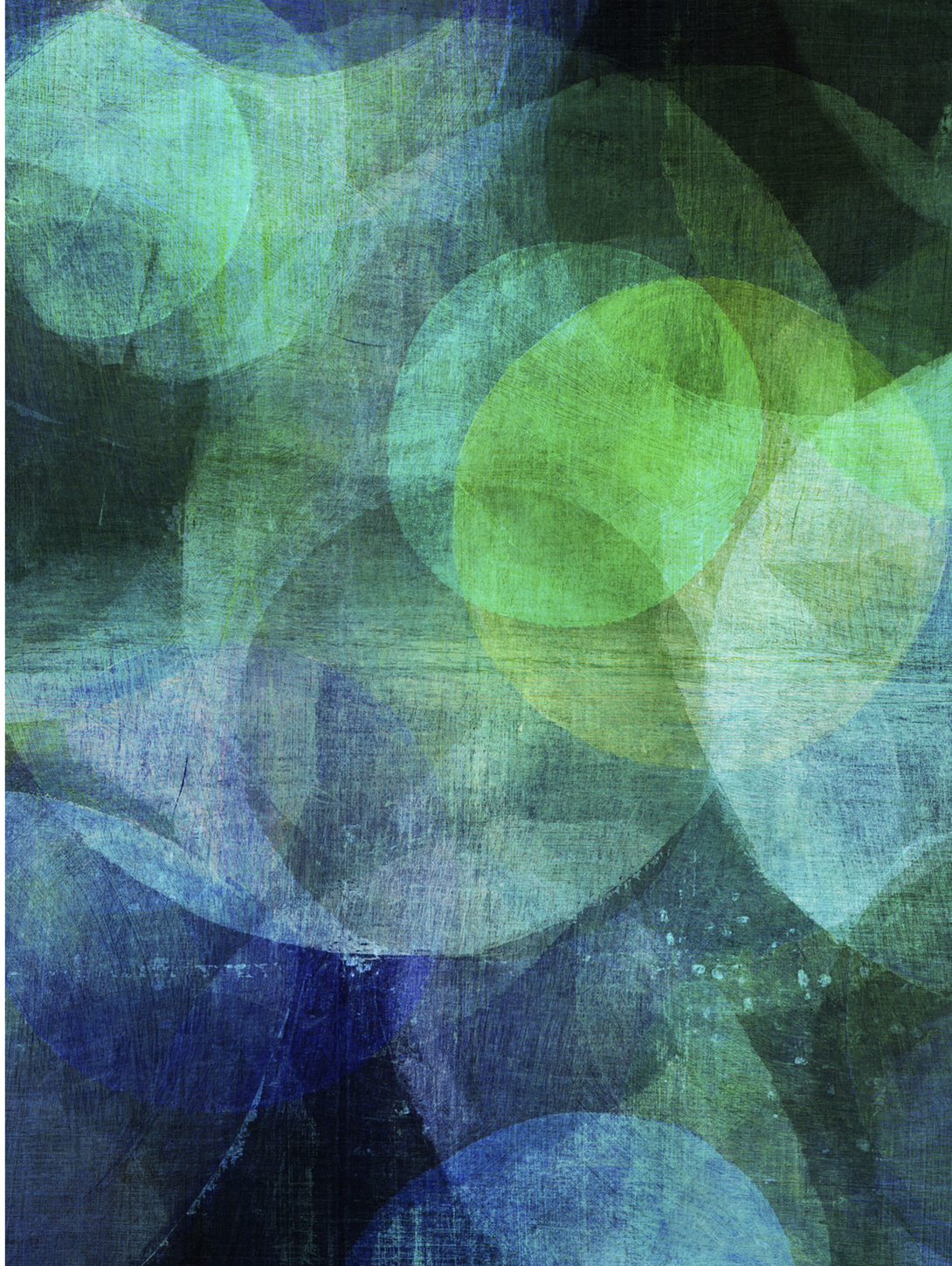
Coherence: How coherent is the summary? How natural is the ordering of the facts? The summary should be well structured and well organized and have a natural ordering of the facts.

Answers

Best: ☐ Worst: ☐

SET-UP

What did we do?



CRITERIA

Three basic criteria, that are given to the raters:

Grammaticality	Is the summary written in well-formed English?
Coherence*	Is the summary well structured and well organized and does it have a natural ordering of the facts?
Conciseness/repetition	Does the summary avoid unnecessary repetition including whole sentences, facts or phrases?

* Note the use of “and” (three times!) that typically indicates complex constructs that could be split up into more basic parts.

MATERIALS

- Summaries by 4 systems + humans (we'll call these "systems")
- 20 summaries per system
- Ratings for all possible combinations of systems = 200 items
- 1800 ratings (200 items * 3 ratings * 3 criteria), converted to HITs

PARTICIPANTS

- MTurk workers from English-speaking countries
- No fixed number of participants! (Between 1 and 200 for each criterion.)

PARTICIPANTS

- MTurk workers from English-speaking countries
- No fixed number of participants! (Between 1 and 200 for each criterion.)

Note:

- We would want to add: probably a trade-off here with #judgments per evaluator.
- 1 rating per evaluator means reliability is hard to assess.

PAYMENT

Adjustment to match minimum wage requirements:

- Original payment: $0.15 per HIT
- Our payment: $0.34 per HIT

Earlier studies report no (Buhrmester et al. 2011) or positive (Litman et al. 2015) effects of higher wages.

QUALITY CONTROL

Minimal control:

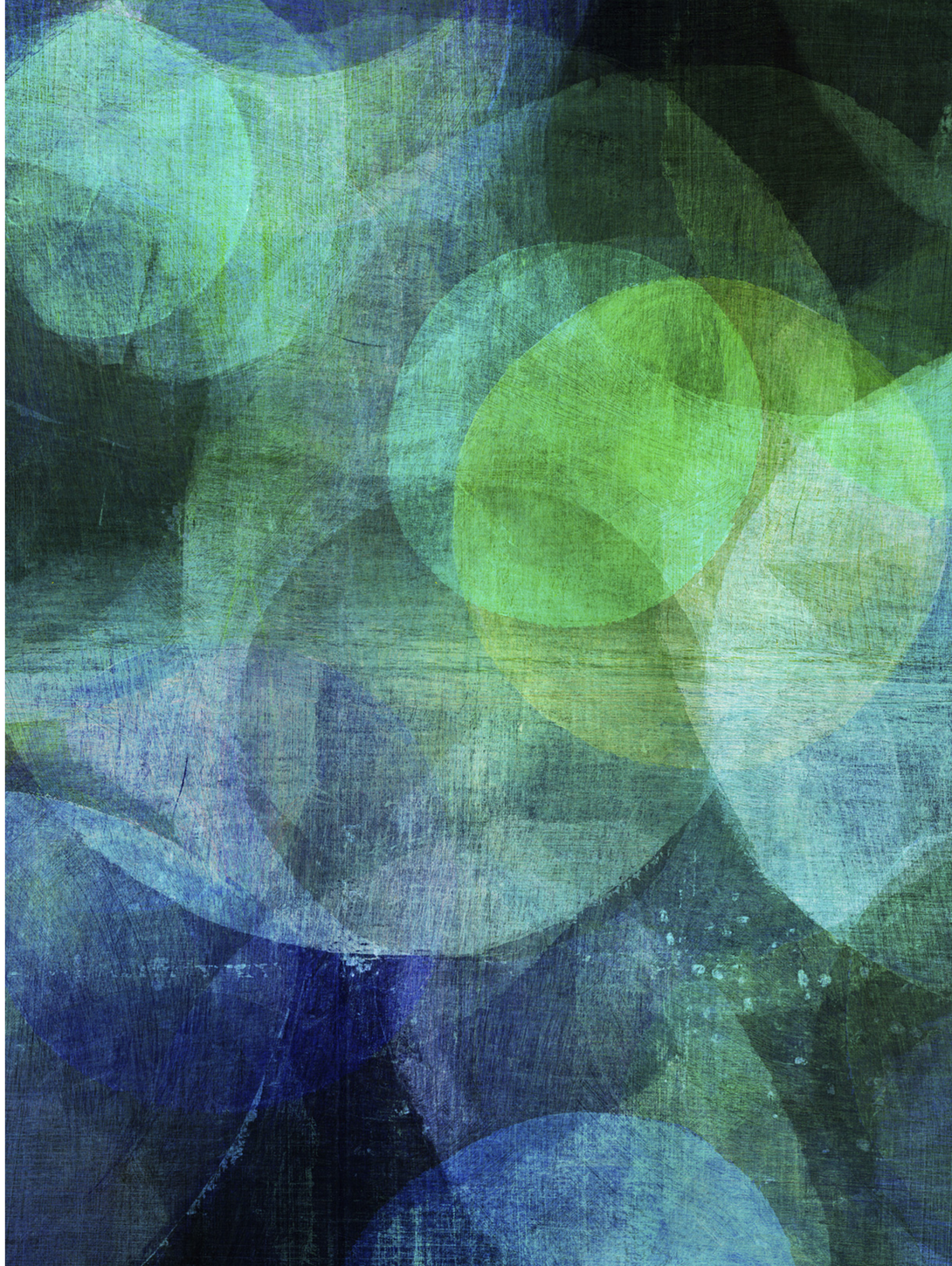
- Inspect results after each batch of HITs.
- Deny future participation to people who failed attention checks.
- Keep all results and hope for the best.

COMPUTING THE SCORES: BEST-WORST SCALING

- A system ‘wins’ when a participant prefers it over another system.
- For each <system, summary, criterion> combination:
  - Score starts at 0.
  - Award a point for every win.
  - Subtract a point for every loss.
  - Scale result to [-100, 100]
- Result: 20 scores per system, per criterion.

RESULTS

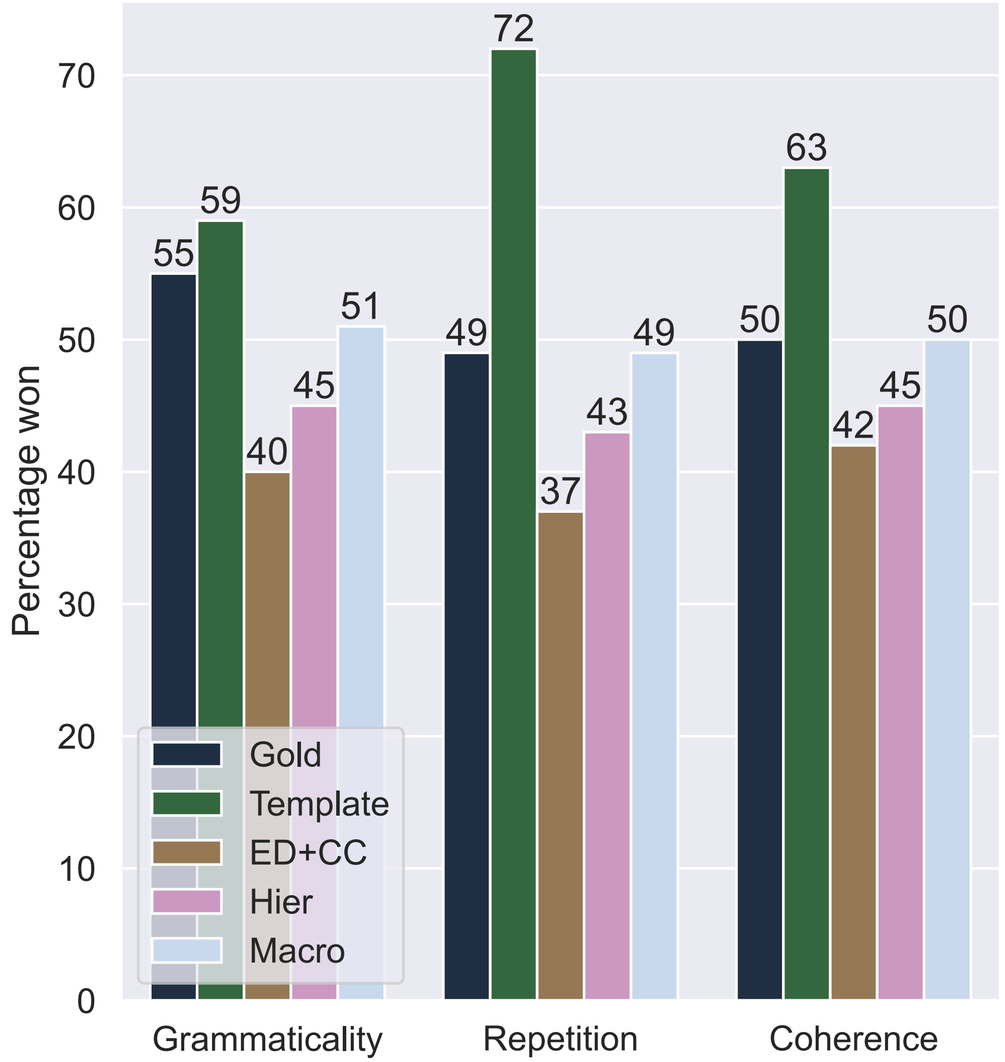
What did we find?



DIFFERENT RESULTS

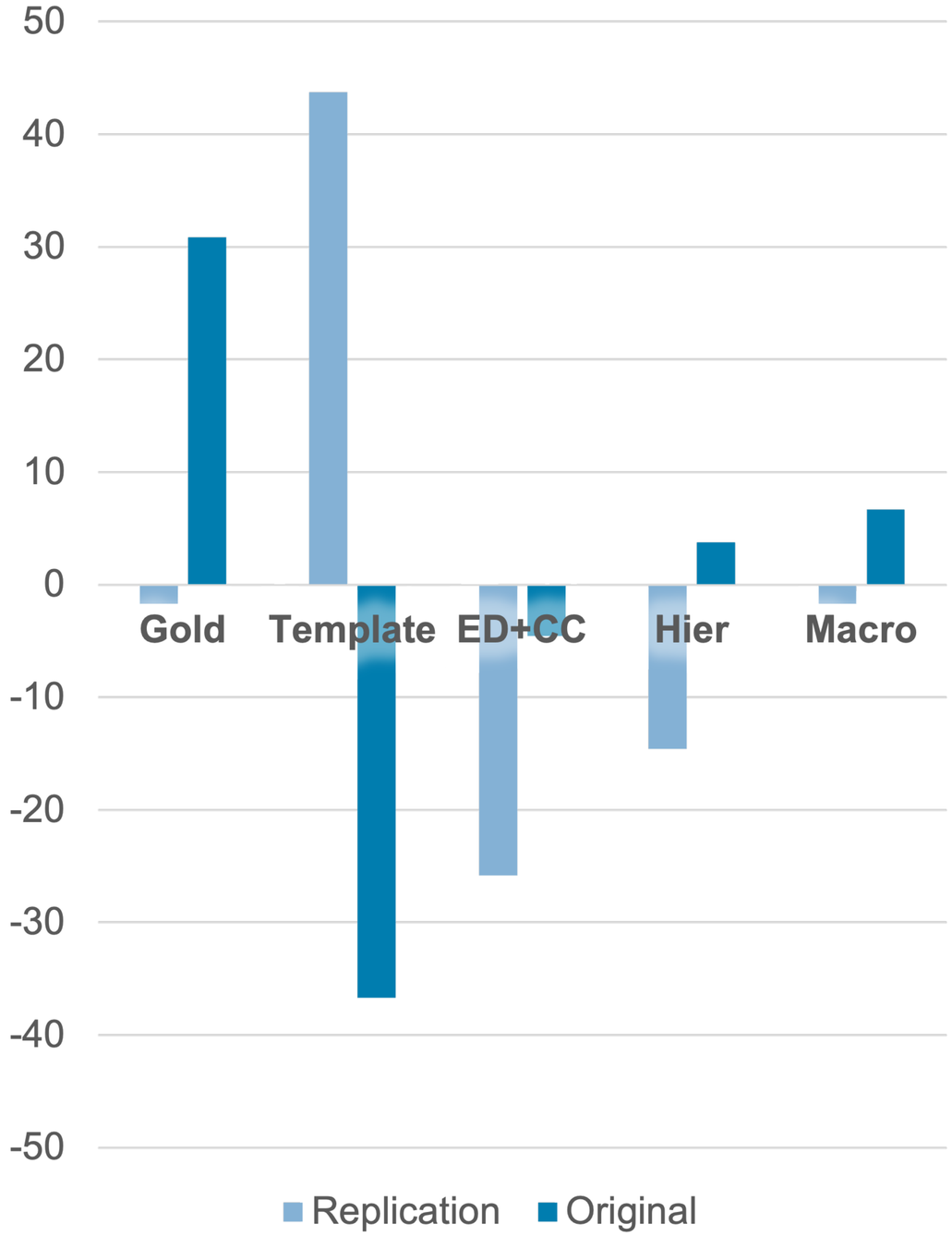
- Original results: “Macro” is best, “Template” worst.
- Our results: “Template” is best.

(No original data available…)



DIFFERENT RESULTS

- Scores look completely different.
- …but this experiment is about *relative scores*.
-



SPEARMAN CORRELATION

- Best-worst scaling is a *relative measure*. So how do the ranks correlate?

Grammaticality

	Original	Ours
Gold	1	2
Template	5	1
Ed+CC	3	5
Hier	2	4
Macro	3	3

$\rho = -0.21$

SPEARMAN CORRELATION

- Best-worst scaling is a *relative measure*. So how do the ranks correlate?

Coherence

	Original	Ours
Gold	1	3
Template	5	1
Ed+CC	4	5
Hier	3	4
Macro	2	2

$\rho = -0.1$

SPEARMAN CORRELATION

- Best-worst scaling is a *relative measure*. So how do the ranks correlate?

Repetition

	Original	Ours
Gold	1	2
Template	5	1
Ed+CC	4	5
Hier	3	4
Macro	2	2

$\rho = -0.05$

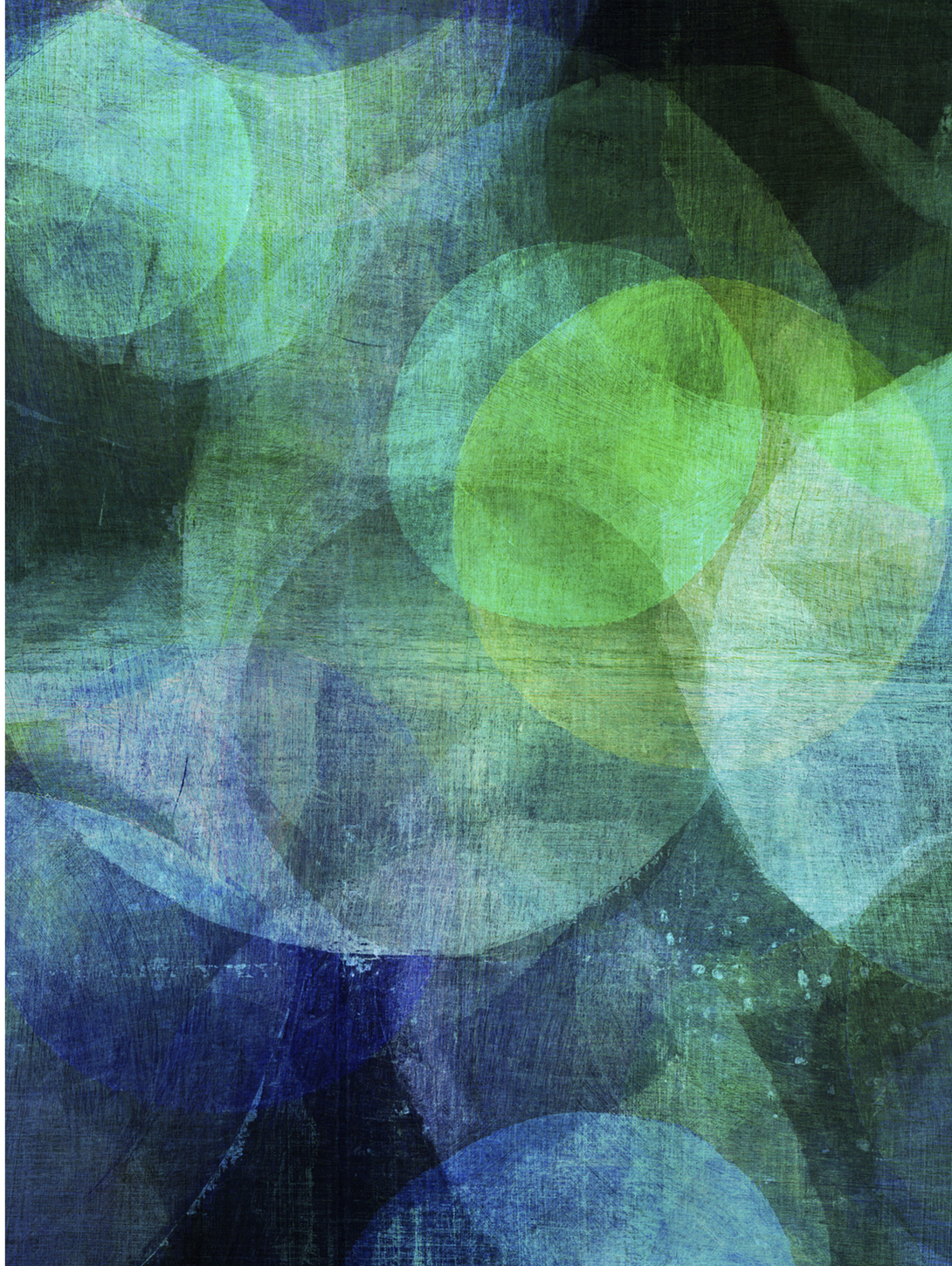
KRIPPENDORFF'S ALPHA

Low inter-annotator agreement:

- Original: 0.47
- Ours: between 0.0438 and 0.203

DISCUSSION

What did we learn?



RELIABLE REPRODUCTION STUDY, OR NOT?

Large amount of noise:

- Inter-annotator agreement is low.
- Very subjective task? Or just low efforts from participants?
- Hard to carry out quality control with small HITs.

Sample size

- Only 20 summaries per system, for each quality criterion.
- Can/should we draw conclusions from such small samples?

SAMPLE SIZE INCREASES?

Earlier work:

- Simonsohn (2015): replications should have 2.5 times the original sample size.
- Van Zwet & Goodman (2022): sample size should depend on effect size, we possibly need up to 16 times the original sample size.

Sample size can be hard to define:

- Number of participants? (Generalising over raters)
- Number of items to rate? (Generalising over items)
- Number of ratings per item? (Determining variability in ratings)

KEY QUESTION

- What do we want to learn from all these reproduction studies?
- Is the current set-up sufficient for this? (Also exact vs. conceptual replications, etc.)

THANK YOU!