

Construction of a benchmark categorization algorithm for Comparative Judgement: a simulation study

San Verhavert, Liesje Coertjens, Sven De Maeyer and Vincent Donche

Paired Comparison

Although comparative judgement (CJ) proves a valid and reliable assessment method, it is characterized by a low efficiency. (Bramley, 2007)

Only a few proposals have been made to increase CJ's efficiency. The most frequently used method consists of some initial Swiss tournament rounds followed by Fisher information based matching (Pollitt, 2012).

Issue 1: Swiss tournament inflates the reliability (Bramley, 2015). The Fisher information based matching might suffer the same problem.

Issue 2: In the context of CJ, there appears a lack of simulation studies testing the efficiency and accuracy of selection algorithms.

Benchmark Categorization

Based on Computerized Categorization Testing (CCT; Spray & Reckase, 1994),

New representations are compared to a set of representations, *close to* one or more benchmarks, on an earlier ranking.

The new representations are classified using a version of SPRT (Reckase, 1983)

Based on Eggen (1999):

- Item selection criteria: Maximum Fisher Information (FI) and Maximum Kullback-Leibler Information (KLI).
- Information calculation reference points
 - For FI:
 - estimated $\hat{\theta}_A$ (a, g)
 - nearest cutting point (NCP; b, h)
 - midpoint closest to $\hat{\theta}_A$ (c, i)
 - For KLI:
 - fixed points (d, j)
 - midpoint of the decision interval, closest to $\hat{\theta}_A$ (e, k)
 - dependent on decision (f, l)
- Stopping criterion: all new representations are classified (variable length; g-l, n) and/or fixed length (Ncomp; a-f [n=100], m [n=500]).

Research questions

RQ1 Which information statistic results in more efficient pair construction and a higher accuracy?

RQ2 Per information statistic, which reference point results in more efficient pair construction, and accuracy?

RQ3 Does SPRT contribute to a higher efficiency and accuracy?

Method

Conditions

12 experimental conditions:

- Ncomp – FI – estimated
- Ncomp – FI – NCP
- Ncomp – FI – midpoint
- Ncomp – KLI – fixed
- Ncomp – KLI – midpoint
- Ncomp – KLI – decision
- SPRT – FI – estimated
- SPRT – FI – NCP
- SPRT – FI – midpoint
- SPRT – KLI – fixed
- SPRT – KLI – midpoint
- SPRT – KLI – decision

2 control conditions using random selection:

- Ncomp – random
- SPRT – random

Simulation details

The ranking from a CJ assessment on argumentative writing was used.

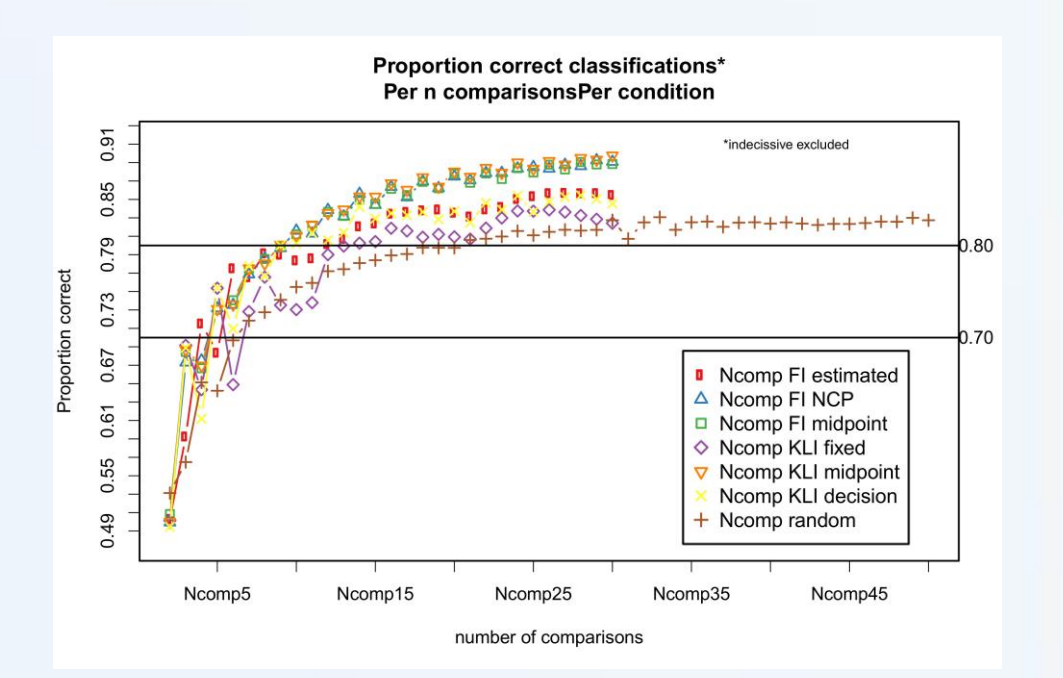
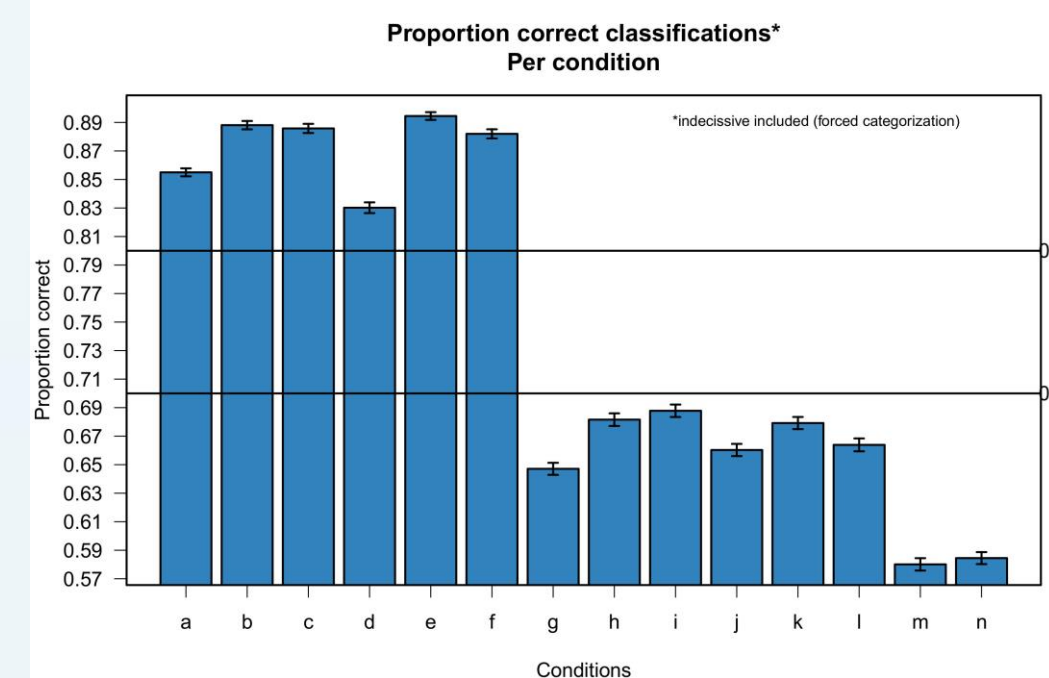
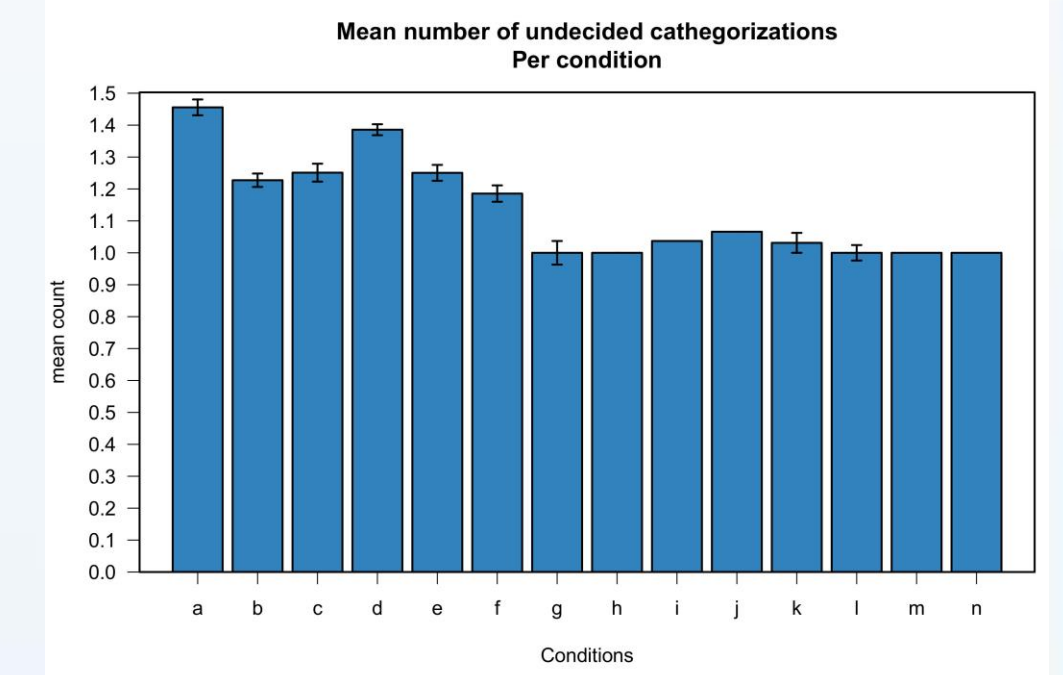
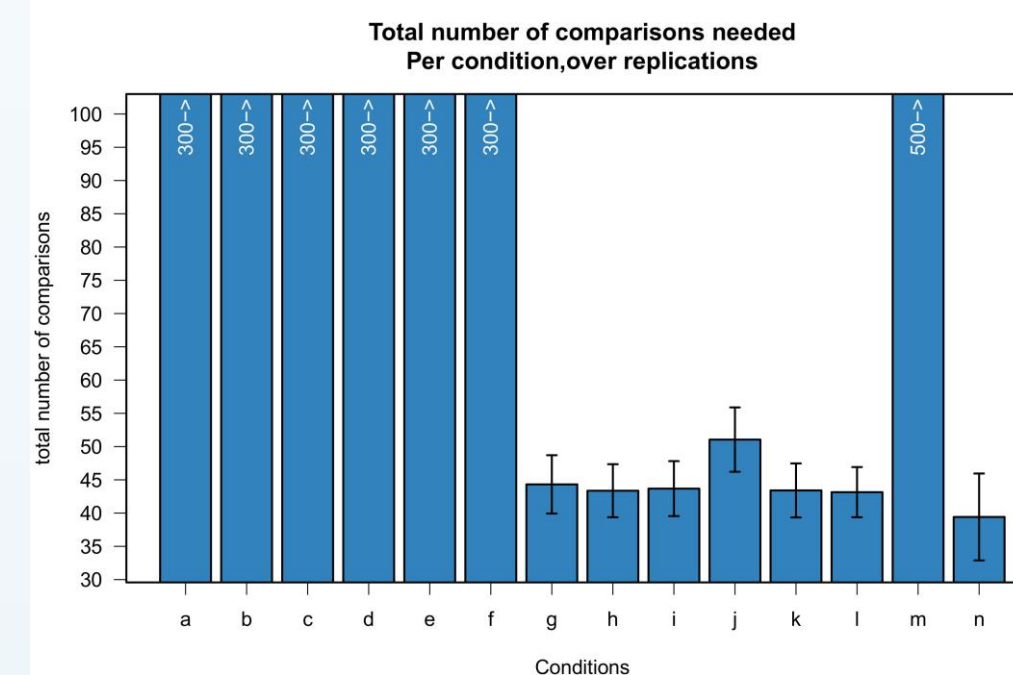
10 representations were randomly selected as new representations.

Benchmarks:

- BM1 at -2.29, and 17 “close to” representations
- BM2 at -0.06, and 15 “close to” representations

SPRT parameters:

- $\delta = 3$
- $\alpha = \beta = 0.05$
- 1000 replications per condition



Results and preliminary conclusion

Information based selection (a-l) requires more comparisons but leads to a higher accuracy than random selection (m-n).

Fixed length CJs (a-f) lead to a much higher accuracy [but more undecided categorizations] compared to variable length CJ (g-l).

Accuracy rises with the number of comparisons. Between 9 and 12 comparisons all selection

algorithms reach a proportion correct of 0,80 In fixed length as in variable length CJ the same four selection criteria appear most accurate:

- Fisher information on nearest cutting point
- Fisher information on the midpoint of the decision interval
- Kullback-Leibler information on the midpoint of the decision interval (less in variable length)
- Kullback-Leibler information on the decision (not in variable length)

Future directions

Applying the algorithm in a real assessment (preliminary results)

Looking into the SPRT parameters in the light of the most efficient and accurate selection criteria.

Testing the limits: how do the most efficient and accurate selection criteria behave under random judgements?

Extending the selection algorithm to more than three categories



www.D-PAC.be



orcid.org/0000-0003-0633-9753

Contact

San Verhavert

San.verhavert@uantwerpen.be

Faculty of Social Sciences – Department

Instructional and Educational Sciences

Researchgroup Edubron

D-PAC project

www.eduBROn.be | www.D-PAC.be

References:

- Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms, *Techniques for monitoring the comparability of examination standards* (1st ed., pp. 246–300). London, U.K.: Qualifications and Curriculum Authority.
- Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgement*. Cambridge Assessment Research Report. Cambridge, U. K.: Cambridge Assessment.
- Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the Sequential Probability Ratio Test. *Applied Psychological Measurement, 23*(3), 249–61.
<http://doi.org/10.1177/01466219922031365>
- Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300.
<http://doi.org/10.1080/0969594X.2012.665354>
- Spray, J. A., & Reckase, M. D. (1994). The selection of test items for decision making with a computer adaptive test. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New Horizons in Testing* (pp. 237–255). New York, NY: Academic Press, Inc.