04/11/2016

Comparative Judgement and Scale Separation Reliability (SSR)

Yes, but what does it mean?

San Verhavert
Sven De Maeyer
Vincent Donche
Liesje Coertjens

http://www.D-PAC.be san.verhavert@uantwerpen.be

2

Overview

- Introduction
- Reliability in theory
- Research Method
- Results and Discussion
- Conclusion

3

Introduction

How can we reduce the number of comparisons in Comparative Judgement (CJ) to arrive to a similar rank order with the same level of reliability?

- Scale Separation Reliability (SSR)
  - Internal consistency
  - Yes, but what does it mean?
- How many comparisons do we need anyway?

Research Questions

- What does the reliability measure mean in CJ?
- When is a CJ assessment reliable (enough) if no adaptive algorithms are used?

D-PAC

Overview

- Introduction
- Reliability in theory
- Research Method
- Results and Discussion
- Conclusion

D-PAC

Reliability in theory

- CTT – G Theory – IRT (Brennan, 2011)

D-PAC

Reliability in theory 7

Scale Separation Reliability: Formula

- classical Rasch (Bramley, 2015)

$$\alpha = \frac{G^2}{(1+G^2)}$$

$$G = \frac{trueSD}{RMSE}$$

$$RMSE = \sqrt{MSE} = \frac{\sum_i SE_i^2}{n}$$

D-PAC

Reliability in theory 8

Scale Separation Reliability: Formula

- classical Rasch (Bramley, 2015)

$$\alpha = \frac{\frac{trueSD^2}{MSE}}{\left(1+\frac{trueSD^2}{MSE}\right)}$$

$$G = \frac{trueSD}{RMSE}$$

$$RMSE = \sqrt{MSE} = \frac{\sum_i SE_i^2}{n}$$

D-PAC

Reliability in theory 9

Scale Separation Reliability: Formula

- classical Rasch (Bramley, 2015)

$$\alpha = \frac{trueSD^2}{obsSD^2}$$

D-PAC

Reliability in theory

10

Scale Separation Reliability

$$\frac{trueSD^2}{obsSD^2}$$

$$\frac{obsSD^2 - MSE}{obsSD^2} = R^2$$

$$trueSD^2 = obsSD^2 - MSE$$

- CTT – G Theory – IRT (Brennan, 2011)
- Traditional view: Variance in observed scores attributable to true scores (Brennan, 2011; Webb, Shavelson & Heartel, 2006)
  - Rasch measurement (Andrich, 1982)
  - Comparative Judgement (e.g. Bramley, 2015)
  - Paired Comparison (Dunn-Rankin, Knezek, Wallace & Zhang, 2004; Gulliksen and Tukey, 1958)
- Model prediction: Scalability (Gulliksen and Tukey, 1958)

D-PAC

Reliability in theory

11

Scale Separation Reliability

- Scalability → CTT

  "[T]he correlation between two sets of observations, taken this correlation is entirely accounted for by the true values." (Gulliksen and Tukey, 1958, p105)

- Correlation between observed and *true values*
  - Correlation between variations of the same assessment (Bramley, 2015)

D-PAC

12

Reliability in theory

- Correlation between variations of the same assessment

1. Internal consistency
   - Split halves
2. Parallel or equivalent forms
   - Different judges – same representations and algorithm
     - Split halves
     - inter-rater
   - Different judges – same comparisons
   - Same judges – same and different representations
3. Test-Retest reliability
   - Same judges – same comparisons

D-PAC

(Bramley, 2015; Webb, Shavelson & Heartel, 2006)

13

Reliability in theory: When?

- (Random allocation/distributed random allocation)

1. Theoretical: $\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2}$
2. Empirical: correlation between intermediate and final ranking (Wheadon, 2015)
   - 10 comparisons per representation

D-PAC

14

Overview

- Introduction
- Reliability in theory
- Research Method
- Results and Discussion
- Conclusion

D-PAC

15

Research method

- Data:
  - 27 datasets of try-outs and data collections in D-PAC
- Analysis:
  - What does reliability mean?

D-PAC

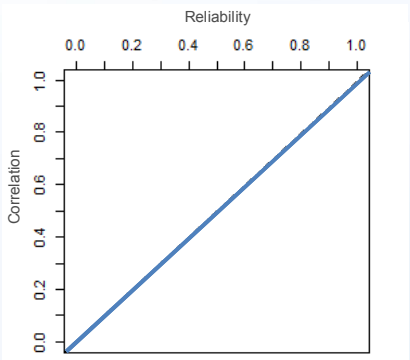
16

Research method

- Data:
  - 27 datasets of try-outs and data collections in D-PAC
- Analysis: What does reliability mean?
  - every possible split half or 1000 random split halves
  - Or Different groups of assessors
  - Pearson's r and Kendall's τ
  - Plot correlation x reliability

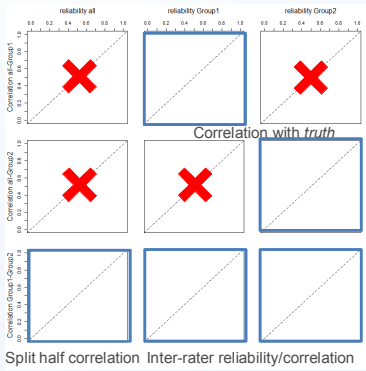
D-PAC

17



D-PAC

18



D-PAC

19

Research method

- Data:
  - 27 datasets of try-outs and data collections in D-PAC
- Analysis:
  - What does reliability mean?
  - When is an assessment reliable?

D-PAC

20

Research method

- Data:
  - 27 datasets of try-outs and data collections in D-PAC
- Analysis: When is an assessment reliable?
  - Calculated assessment rounds
    - 1 round = 1 comparison per representation
    - Issues

D-PAC

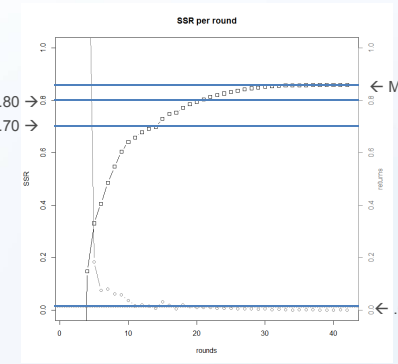
21

Research method

- Data:
  - 27 datasets of try-outs and data collections in D-PAC
- Analysis: When is an assessment reliable?
  - Calculated the rounds
  - Estimated the ranking and the reliability after each round
  - Calculated the returns: difference in reliability between rounds
  - Plot Rounds x Reliability

D-PAC

22



D-PAC

23

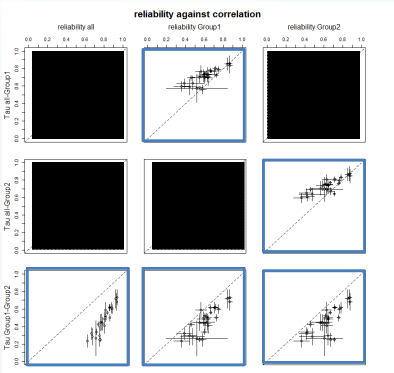
Overview

- Introduction
- Reliability in theory
- Research Method
- Results and Discussion
- Conclusion

D-PAC

24

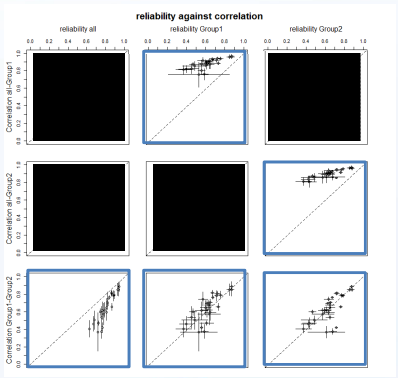
Kendall's τ correlation



D-PAC

Pearson's r correlation

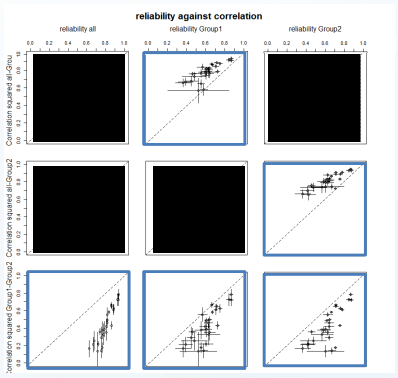
25



D-PAC

R^2

26



D-PAC

Results and discussion

27

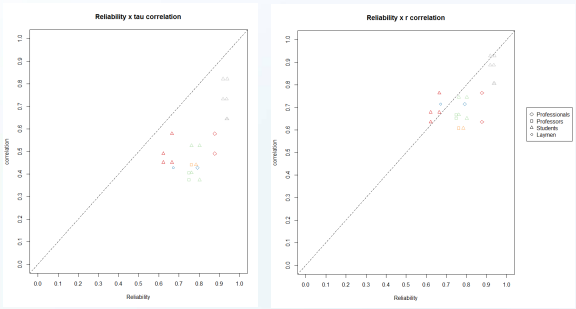
What?

- Kendall's τ correlation
  - Correlation with *truth*
- Pearson's r correlation
  - Inter-rater reliability
- R^2
  - Correlation with *truth*

- Confirmation: true inter-rater?

D-PAC

28



D-PAC

Results and discussion

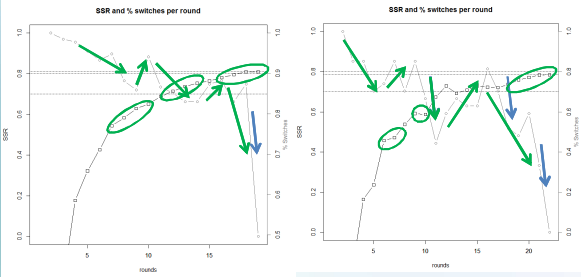
29

What?

- Kendall's τ correlation
  - Correlation with *truth*
- Pearson's r correlation
  - Inter-rater reliability
- R^2
  - Correlation with *truth*
- Confirmation: true inter-rater?
- Alternative internal consistency: Stability

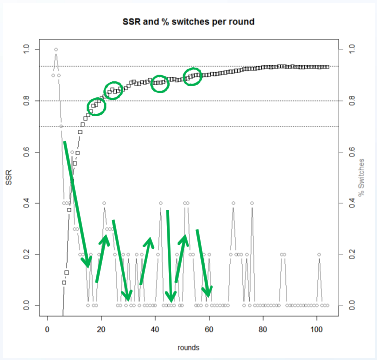
D-PAC

30



D-PAC

31



D-PAC

Results and discussion

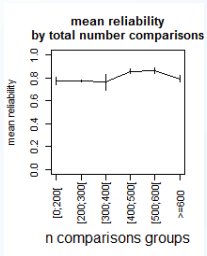
32

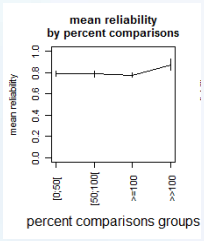
What? – Differences

- Main effects

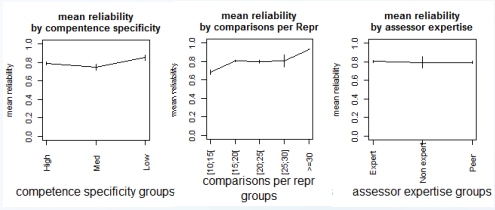
D-PAC

33



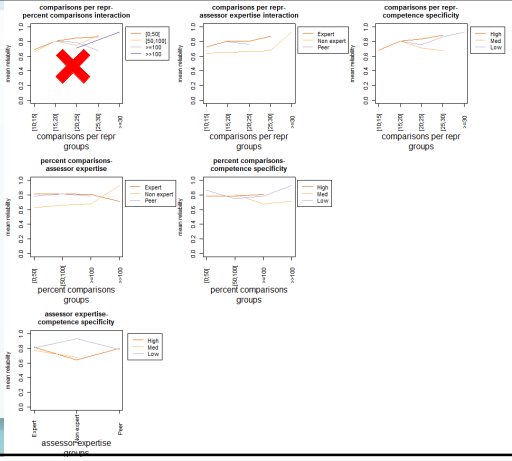


D-PAC



What? – Differences

- Main effects
- Interaction effects



Results and discussion 37

When?

- Asymptote: 14 comparisons[1]
  - Min: 14; Mean: **20**; Max: 40
  - Exception
- SSR = .70: 9 comparisons[1]
  - Min: 9; Mean: 12; Max: 20
- SSR = .80: 13 comparisons[1]
  - Min: 13; Mean: 17; Max: 25

[1]per representation

D-PAC

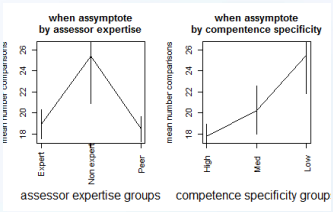
Results and discussion 38

When? – Differences

- Main effects

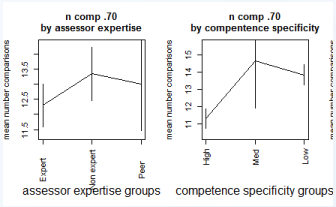
D-PAC

39

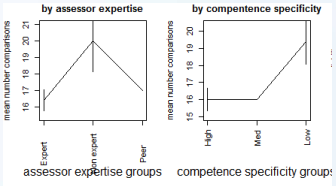


D-PAC

40



41



Results and discussion

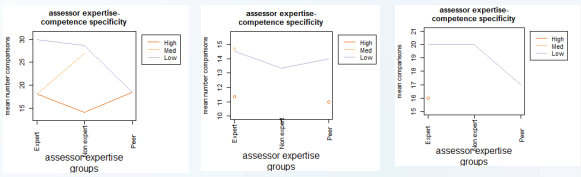
42

When? – Differences

- Main effects
- Interaction effects

43

Asymptote .70 .80



D-PAC

Results and discussion

44

When? – Differences

- Main effects
- Interaction effects
- Feedback

D-PAC

Results and discussion

45

When? – Differences

- Feedback
  - Asymptote
    1. No feedback
    2. Comparative feedback
    3. Pro's – Con's
  - .70
    1. Comparative feedback
    2. No feedback
    3. Pro's – Con's
  - .80
    1. Comparative feedback
    2. No feedback
    3. Pro's – Con's

D-PAC

When? – Differences

- Main effects
- Interaction effects
- Feedback
- Training
  - Asymptote No training is faster
  - .80: only training group

D-PAC

Overview

- IntroductionReliability in theory
- Research Method
- Results and Discussion
- Conclusion

D-PAC

Conclusion

- SSR
  - Classical reliability
  - Inter-rater reliability (?)
  - Stability of ranking
    - Internal consistency
  - Depends on
    - number of comparisons per representation
    - Specificity
    - Expertise

D-PAC

49

Conclusion

- Reliable between 9 and 20 comparisons
  - Depends on
    - Specificity
    - Expertise
    - Feedback (type)
    - Training (??)

D-PAC

50

Conclusion

- Limitations
  - Practice oriented: Exploratory
  - Inter-rater and not independent groups
  - Limited number of assessments

D-PAC

51

References

Andrich, D. (1982). Index of Person Separation in Latent Trait Theory, the Traditional KR-20 Index, and the Guttman Scale Response Pattern. *Education Research and Perspectives*, *9*(1), 95–104.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgement* (Cambridge assessment research report). Retrieved from www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf

Brennan, R. L. (2011). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Dunn-Rankin, P., Knezek, G. A., Wallace, S. R., & Zhang, S. (2004). *Scaling Methods* (2 edition). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gulliksen, H., & Tukey, J. W. (1958). Reliability for the law of comparative judgment. *Psychometrika*, *23*(2), 95–110. https://doi.org/10.1007/BF02289008

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 Reliability Coefficients and Generalizability Theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 81–124). Amsterdam, The Netherlands: Elsevier.

Wheadon, C. (2015, February 10). The opposite of adaptivity [Blog post]. Retrieved from https://www.nomoremarking.com/blog/kkfsnY5GzeoFscAcH

D-PAC


52
A person with one watch knows what time it is; a person with two watches is never quite sure!
Edubron
Universiteit Antwerpen
UNIVERSITEIT GENT
iwt
iMinds
D-PAC