20/03/2015

Benchmark Categorization with Comparative Judgement

San Verhavert

D-PAC

san.verhavert@uantwerpen.be

Digital Platform for the Assessment of Competences

Competencies and rubrics

- Marking and rubrics not necessarily valid and reliable judgements (Johnson & Svingby, 2007; Heldsinger & Humphry, 2010)
- Judgments are relative, not absolute (Lamming,1990)

=>lower accuracy in rubrics

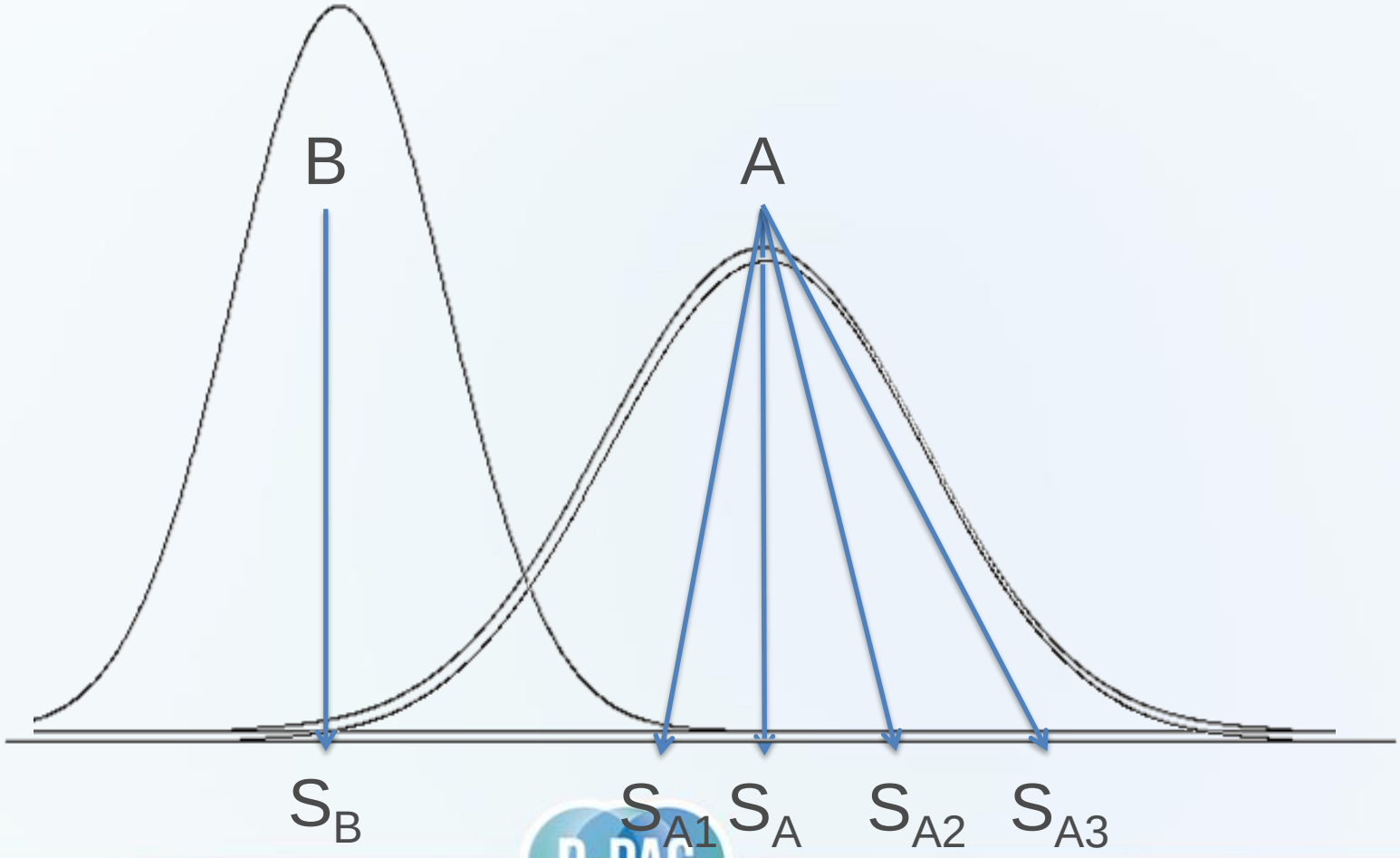
- To reductionist (Pollitt, 2004)

Solution?

- Bramley, Bel and Pollitt (1998)
- The Law of Comparative Judgement (Thurstone, 1927)

(based on Bramley, 2007)

The Law of Comparative Judgement
(Thurstone, 1927)

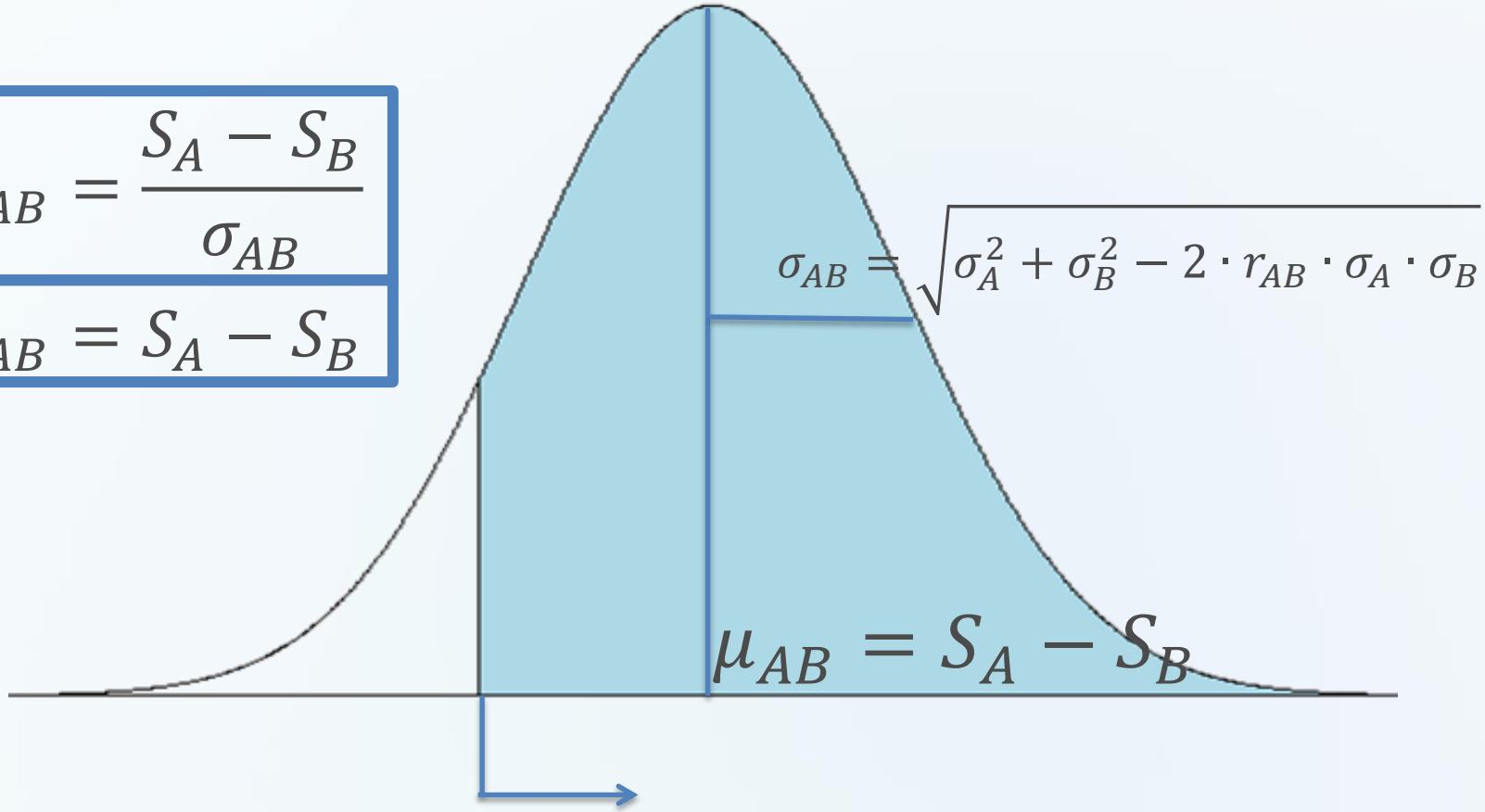


(based on Bramley, 2007)

The Law of Comparative Judgement
(Thurstone, 1927)

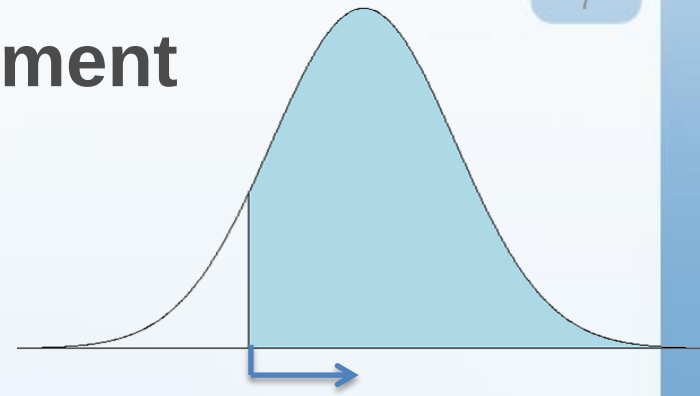
$$X_{AB} = \frac{S_A - S_B}{\sigma_{AB}}$$

$$X_{AB} = S_A - S_B$$



(based on Bramley, 2007)

The Law of Comparative Judgement
(Thurstone, 1927)



$$X_{AB} = S_A - S_B$$

$$p(\text{'A beats B'}) = p(A > B) = \frac{\exp(S_A - S_B)}{1 + \exp(S_A - S_B)}$$

Bradley-Terry-Luce Model (BTL; Bradley & Terry, 1952; Luce,1959)

(based on Bramley, 2007)

Rasch model
(Rasch, 1960)

$$p(\alpha_{vj} = 1 | \alpha_j, \tau_v) = \frac{\exp(\alpha_j - \tau_v)}{1 + \exp(\alpha_j - \tau_v)}$$

$$p(A > B) = \frac{\exp(S_A - S_B)}{1 + \exp(S_A - S_B)}$$

$$\Leftrightarrow \mathrm{p}(x_{AB} = 1 | \alpha_A, \alpha_B) = \frac{\exp(\alpha_A - \alpha_B)}{1 + \exp(\alpha_A - \alpha_B)} \quad \text{(BTL)}$$

$$\forall\, x_{AB} \epsilon' comparison\ outcome' : A > B \ \Leftrightarrow x_{AB} = 1$$

(Andrich, 1978)

Efficiency?

- Time cost and monotony of the task (e.g. Bramley, 2007)
- 1224 judgement, 135 representations, 55 judges
  → Alpha= .68
- Solution: Adaptive selection (Pollitt, 2004, 2012)
  - Derived from CAT
  - Preliminary estimates to select pairs
  - >50% comparisons less
  - BUT

Adaptive selection and reliability

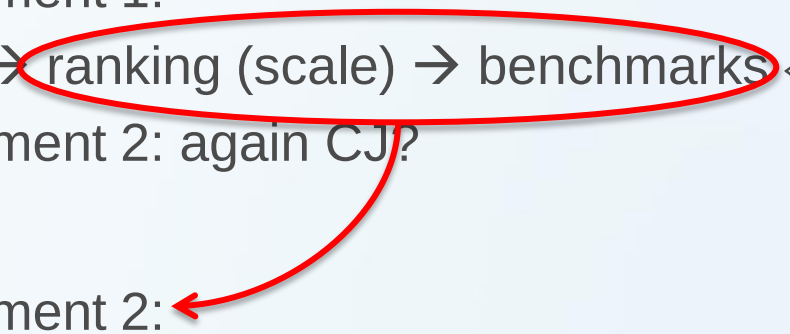
- Inflation of alpha
- Independence of judgments violated
- Circularity issue

- Robustness of BTL?

The solution in D-PAC

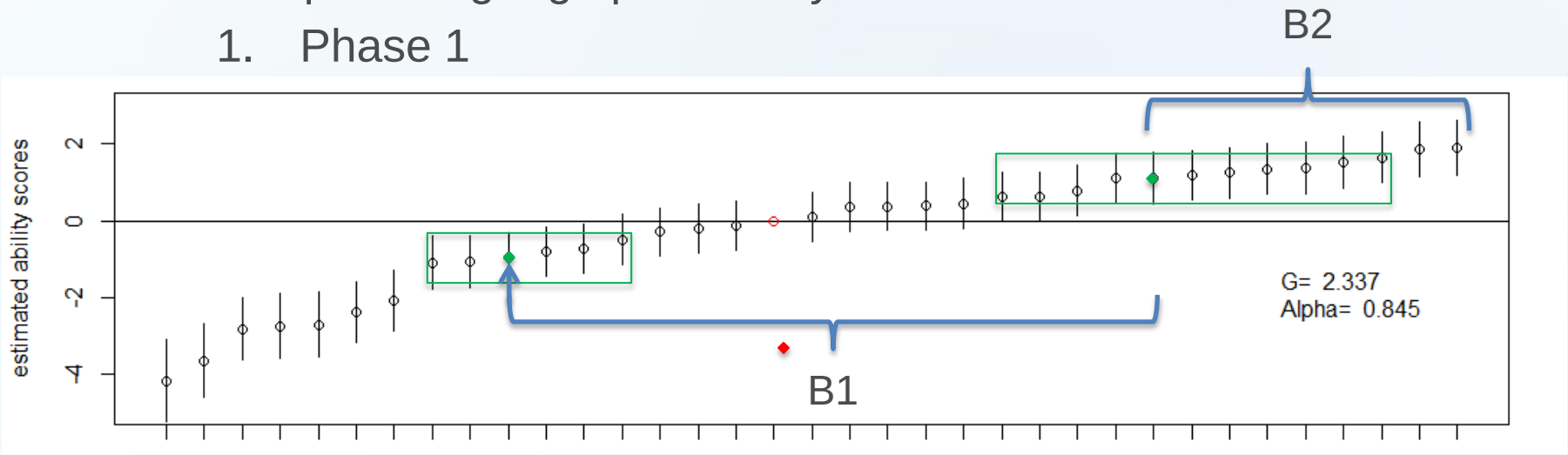
Benchmark categorization

- Problem: categorization
  - ↳ Assessment 1:
    - ↳CJ → ranking (scale) → benchmarks ⇐ Time consuming
  - ↳ Assessment 2: again CJ?
- Solution:
  - ↳ Assessment 2:



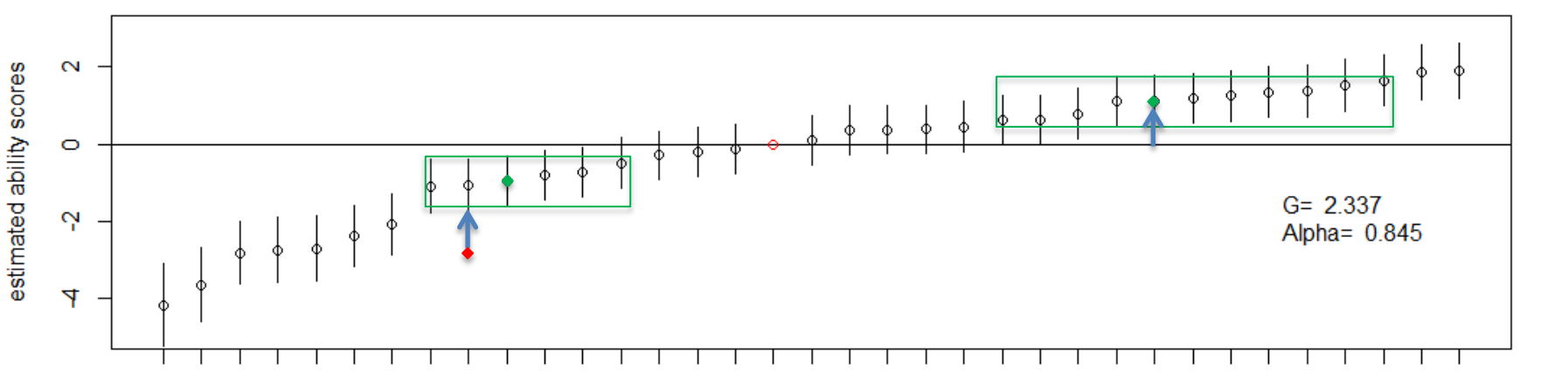
Benchmark categorization

- How do I accurately determine in which category a script belongs?
  - Example: Language proficiency
    1. Phase 1



Benchmark categorization

- How do I accurately determine in which category a script belongs?
  - Example: Language proficiency
    2. Phase 1



Benchmark categorization

- How do I accurately determine in which category a script belongs?
  - Example: Language proficiency
    3. Phase
       - Previous judgements → $\hat{\alpha}_i$
       - Nearest cutting point:

         $\min_c(|\alpha_c - \hat{\alpha}_i|)$ with c={1,2} and i={1,…,n}
       - Maximum Fischer Information (MFI):

         $\max_j\left(I_j(\alpha_c) = p_j(\alpha_c)\left(1 - p_j(\alpha_c)\right)\right)$ with j={1,…, m}

       and c= chosen cutting point

$$p(\alpha_i) = \frac{\exp(\alpha_i - \alpha_j)}{1+\exp(\alpha_i - \alpha_j)} \text{ (BTL)}$$

$$p(\alpha_i) = \frac{\exp(\alpha_i - \tau_j)}{1+\exp(\alpha_i - \tau_j)} \text{ (Rasch)}$$

(Eggen and Straetmans, 2000)

Benchmark categorization

- How do I accurately determine in which category a script belongs?
  - Example: Language proficiency
    3. Phase
       - …
       - Stopping criterion:
         - # judgements per representation=10
         - sequential probability ratio test (SPRT) (Wald, 1947)

(Eggen and Straetmans, 2000)

Simulation study

- Benchmark categorization vs semi-random CJ
  - SPRT
  - Item selection (Eggen and Straetmans, 2000)
    1. MFI at current $\hat{\alpha}_i$
    2. MFI at midpoint critical inequality interval
    3. Kullback-Leibler informations at nearest cutting point
- Ability estimates (55 judges, 135 papers, 1224 judgements)
- 10 replications per condition
- Measures:
  - # comparisons
  - χ^2
  - cohen's κ

Thank You!

Questions?

Edubron Universiteit Antwerpen

UNIVERSITEIT GENT

iMinds ILAB.O

D-PAC