# Response shift analysis at the item level among multiple groups using Rasch models: a simulation study

Myriam Blanchin, Odile Stahl, Yseulys Dubuy, Véronique Sébille

# Response shift analysis at the item level among multiple groups using Rasch models: a simulation study

Myriam Blanchin[1*], Odile Stahl[1], Yseulys Dubuy[1], Véronique Sébille[1]

[1] *Nantes Université, Université de Tours, INSERM, UMR1246 SPHERE "methodS in Patient-centered outcomes and HEalth ResEarch", Nantes, France*

Contact: myriam.blanchin@univ-nantes.fr

## Introduction

Integrating covariates in item-level response shift (RS) analysis could help investigating whether clinical (pathology, treatment…) or psychological (anxiety, perceived stress,...) characteristics of patients are associated with RS and changes in Patient-Reported Outcomes (PRO). However, few methods are available to analyze RS between more than two groups of patients, in particular when using Rasch Measurement Theory (RMT).

**Objectives**: A simulation study was performed to evaluate the performance of a method for item-level RS analysis using RMT among multiple groups of patients.

## Methods: Data simulation

Longitudinal Partial-Credit model
2 times of measurement (T1;T2)
3 groups of patients: G0 (reference group), G1 & G2 (focal groups)

**SIMULATION PARAMETERS**

- Sample size (N): 200 or 300 (balanced groups)
- Number of items (J): 4 or 7
- Number of response categories (M): 4 or 7
- Normal distribution of the latent variable:
  - No group effect or positive group effect (G2>G1>G0)
  - No time effect or positive time effect (same PRO improvement for G1 & G2)

**RS SCENARIOS**

- No RS
- RS affected one focal group (G2)
- RS affected both focal groups (G1 & G2)
  - Similarly: same RS effects
  - Differentially: different magnitudes of RS effects (G2>G1)
- RS affected all groups the same way

**TYPE OF RS**

One type per scenario, one affected item

- **Uniform recalibration (UR):**

Same shift of all item difficulties of one item

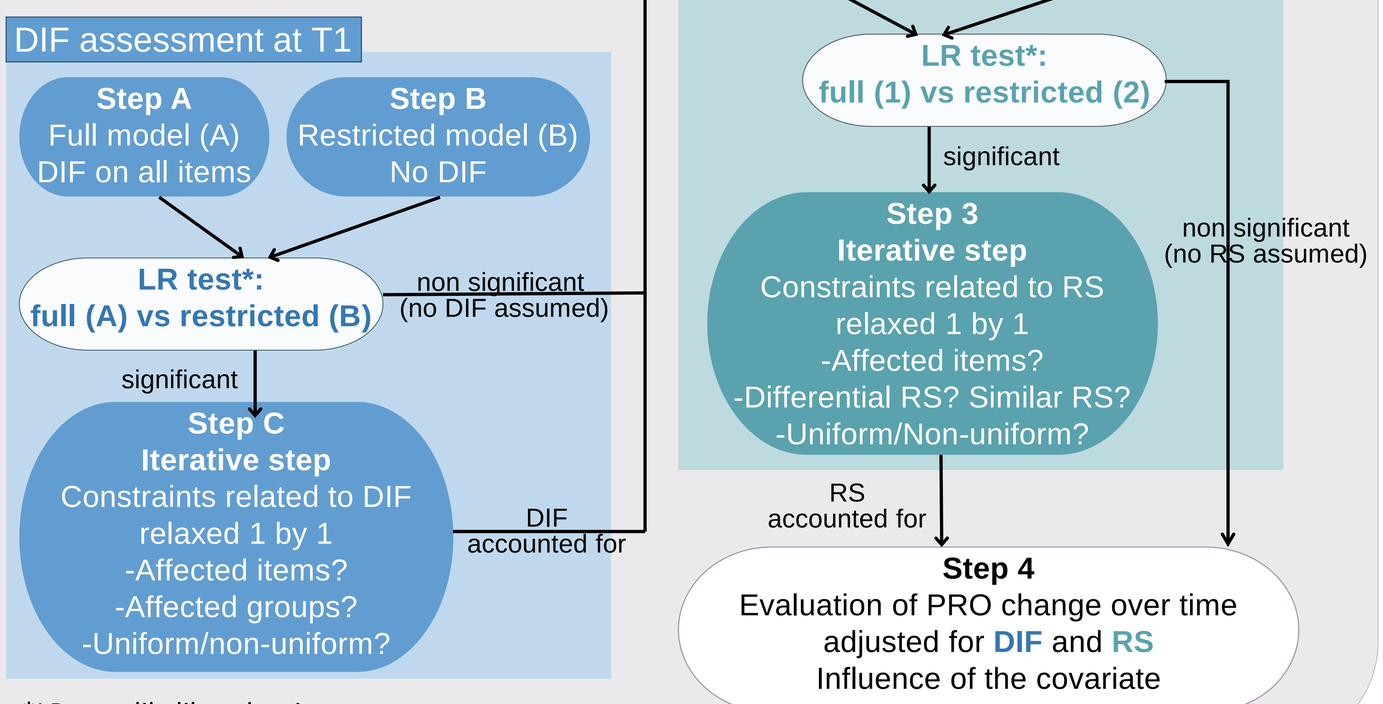
- **Non-uniform recalibration (NUR):**

Various shifts on some item difficulties of one item

336 combinations - 500 replications

## Methods: Data analysis replicated on each simulated dataset

"RespOnse Shift ALgorithm at the Item level"
**ROSALI** (Hammas et al., 2020)
Stata module –rosali- on ssc
Freely accessible on **https://pro-online.net**
Extended to three groups of patients




*LR test: likelihood ratio test

## Methods: Performance criteria

**Criteria assessed at the end of ROSALI**

**Rate of DIF detection**
Proportion of datasets for which DIF effects were detected and accounted for
→ Expected to be close to 5% as no DIF was simulated in any scenario

**Rate of RS detection**
Proportion of datasets for which RS effects were detected and accounted for
→ Expected to be close to 5% when RS was not simulated (false detection)
→ Expected to be high when RS was simulated

## False detection of DIF and RS

- All rates are lower than 3%
- Higher rates of false detection based on LR Tests: [3.6%-7.8%] for DIF, [7.2%-10.8%] for RS

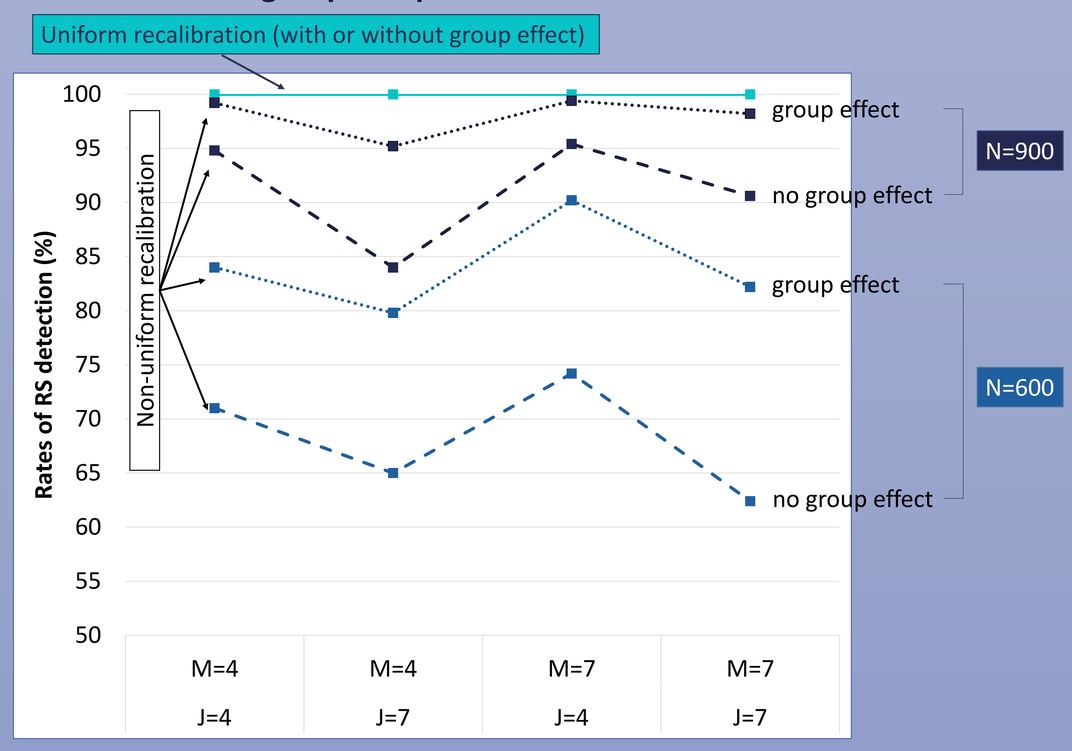
→ Bonferroni correction applied in step C and step 3
→ part of false detection corrected: lower rates of false detection at the end of ROSALI (table opposite)

**Ranges of rates of DIF and RS detection**
No DIF and RS simulated

| N | J | M | DIF | RS |
|---|---|---|---|---|
| 600 | 4 | 4 | 1.8-2.8% | 1.0-2.2% |
| 900 | 4 | 4 | 1.4-2.4% | 0.4-2.4% |
| 600 | 7 | 4 | 0.8-2.4% | 0.6-2.4% |
| 900 | 7 | 4 | 0.8-2.0% | 1.4-2.6% |
| 600 | 4 | 7 | 1.2-2.0% | 0.8-1.6% |
| 900 | 4 | 7 | 1.2-2.4% | 1.4-2.6% |
| 600 | 7 | 7 | 1.4-2.2% | 0.2-2.0% |
| 900 | 7 | 7 | 0.6-1.8% | 0.6-2.2% |

## Example of scenarios with simulated RS

**RS simulated in one group – Improvement of the PRO over time**




- RS detected in 100% of datasets if RS is uniform (whatever the values of the other simulated parameters)
- RS detection decreases as J increases
- RS detection increases as N or M increases or when a group effect has been simulated

## Discussion

Low rates of false detection of DIF and RS

Scenarios with simulated RS
- Uniform recalibration: good performances
- Non-uniform recalibration
  Best performances: RS affecting all groups similarly
  Worst performances: RS affecting one group or two groups in a different way

A total sample size of 600 patients (300 per group) may be insufficient in some contexts to reach a reasonable rate of detection

Iterative steps C and 3: adjustment of pvalues for multiple testing is valuable

**The low rates of false detection indicates that ROSALI satisfactorily prevents from concluding erroneously to DIF or RS. The different patterns of RS inform about the required sample size to correctly detect RS depending on the expected number of affected groups.**

This project has received funding from the European Union's Horizon 2020 Research and Innovation Program under grant agreement No 847782.
Computations performed thanks to the *Centre de Calcul Intensif des Pays de la Loire* (CCIPL) resources.

https://hap2-project.com/