Applying the Generalized Linear Mixed Models to DIF detection

Marina Esposito-Farèse, Christophe Lalanne

Centre international d'études pédagogiques

1, av. Léon Journault, F-92318 Sèvres cedex, France

Background

When constructing a test, we should ensure that the meaning of the items is the same for individuals of all groups: Tests have to be ***equitable***.
The **Rasch model** formulation in terms of **Generalized Linear Mixed Models** (GLMM) is used to detect **differential item functioning** (DIF). This appoach permits the use of the entire sample population to detect DIF items before the introduction of inevitable sources of variability arising from both separate parameter estimations and equating.
We present the results of a comparison study based on simulations (100 replications for 9 different settings). The end point of the study is to establish differences between DIF detection methods.

Comparison of four methods

- CTT
  - Mantel-Haenszel (MH)
  - Mantel-Haenszel with ETS classification
  - Standardization's item discrepancy index

  - DIF assesses how an item functions after matching groups on the construct (ability) measured by the test
  - Need of comparable groups → stratification of the population
  - We work with an observable variable: the total score
  - Definitions depend on proportion-correct

- IRT → Rasch model
  Generalized linear mixed models (GLMM)

  - DIF: **Item characteristic curves** are *significantly* different (Lord 1980)
  - non-observable (latent) variable
  - Probability of response correct regarding the group membership

    $$P(X = x|\theta, Z = z) = P(X = x|\theta), \quad z \in Z$$

    which assumes a measurement model that holds for all $z \in Z$
  - **Inferential statistics**

ETS classification of DIF for Mantel-Haenszel

Scaling the MH common odds ratio $\widehat{\alpha_{MH}}$ giving the effect size: into $MH\ D\text{-}DIF = -2.35 \log(\widehat{\alpha_{MH}})$

Class A $|MH\ D\text{-}DIF| < 1$ **or** $MH\ D\text{-}DIF$ is not statistically different from 0
Class C $|MH\ D\text{-}DIF| > 1.5$ **and** $|MH\ D\text{-}DIF| > 1$ statistically
Class B All others

Standardization's item discrepancy index

$$STD\ P\text{-}DIF = \sum_i w_i(p_{fi} - p_{ri}) / \sum_i w_i \text{ at score level } i$$

$p_{ri} = a_i/(a_i + b_i)$, $p_{fi} = d_i/(c_i + d_i)$ and $w_i = N_{fi}$ subjects at the score level i in the focal group.
The *STD P-DIF* index ranges from -1 to 1

IRT - Rasch model

DIF → effect of the *item* × *group* interaction term on the GLMM formulation. For the subject p and the item i

$$P(X_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1 + \exp(\theta_p - \beta_i)}$$

$$\text{logit}[P(X_{pi} = 1)] = \log\left[\frac{P(X_{pi} = 1)}{P(X_{pi} = 0)}\right] = \theta_p - \beta_i$$

The Rasch model in GLMM matrix formulation is a **random-intercept model** where the person part θ_p is the random effect satisfying $\theta_p \sim \mathcal{N}(0, \sigma^2_\theta)$

$$\eta_{pi} = \sum_{k=0}^{K} \beta_k X_{ik} + \theta_p$$

θ provides a measurement of the latent trait

Modeling DIF
If the item i displaying DIF then

$$\text{logit}[P(X_{pi} = 1)] = \theta_p - (\beta_i + \delta_i)$$

and as $W_{ipk} = X_{ik}Z_{pik}$ is the DIF predictor.

$$\eta_{pi} = \theta_p + \sum_{k=0}(\beta_k X_{ik} + \delta_k X_{ik} Z_{pik})$$

Parameter estimation
GLMM use Marginal maximum likelihood: person parameters are sampled from a distribution so that only their parameters (and not the individual person parameters) enter the likelihood

Effect size measure
- reference $Z = 1$ • focal $Z = 0$

$$-\delta_i = \eta_{pi}|_{Z=1} - \eta_{pi}|_{Z=0}$$

$\exp(-\delta_i)$ is the OR for the focal group vs. the reference group corrected on θ_p

Methodology

Problem: When constructing different tests . . .
What is the chance we have for detecting DIF items? – situation on **TCF®** pretests

	β_1	β_2	...	β_n
s_1	0	1	...	0
s_2	0	1	...	1
⋮	⋮	⋮	⋮	⋮
s_{m-1}	0	1	...	1
s_m	1	0	...	1

	γ_1	γ_2	...	γ_n
t_1	1	0	...	0
t_2	0	1	...	1
⋮	⋮	⋮	⋮	⋮
t_{m-1}	0	1	...	1
t_m	0	1	...	0

...

	δ_1	δ_2	...	δ_n
u_1	1	0	...	0
u_2	0	1	...	1
⋮	⋮	⋮	⋮	⋮
u_{m-1}	0	1	...	1
u_m	1	0	...	1

- 9 settings (*subjects* × *items*)
  300×10 500×10 1000×10
  300×20 500×20 1000×20
  300×40 500×40 1000×40
- 100 replications for each setting

- different tests
- different subjects

- reference and focal groups
- some items may present DIF

- Difficulties are taken randomly (and uniformly) between -2.5 and 2.5
- Tests generation with the R fonction rvmlogic(ltm) following the Rasch model
- Two groups take the test, reference and focal
- For the focal group 20% of the items have DIF
- DIF are taken randomly (and uniformly) from the values { 0.3, 0.5, 0.7 }.
- Position of DIF items is randomly chosen
- Populations are normal, $\theta \sim \mathcal{N}(0,1)$

Remarks

- The DIF introduced is systematic, thus it affects all ability groups equally
- If δ is the DIF, $\hat{\beta}_i$ the theoretical difficulty and β_i the estimated difficulty, then

  $$E(\beta_i + \delta) = \hat{\beta}_i + \delta$$

- It is possible to calculate confidence intervals of the expected difficulty values since *means* are normally distributed

Flagging DIF items

- Mantel-Haenszel 95% CI OR and significant *p*-values
- Mantel-Haenszel as used by ETS 95% CI OR and significant *p*-values
- Standardization: no test is associated. 5% and 10% of the *STD P-DIF* index
- GLMM : 95% CI OR of the estimated effect-size of the interaction

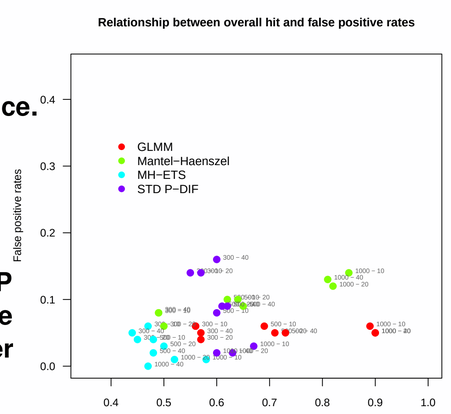
Reporting results

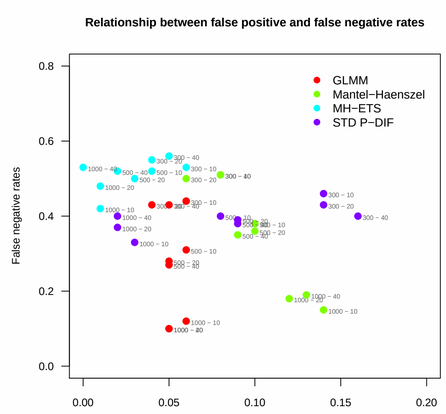
- Hit rates (HR) for each of the DIF introduced, 0.3, 0.5, 0.7 logit
- Overall hit rate – all DIF together
- Overall false positive (FP) & false negative (FN) rates

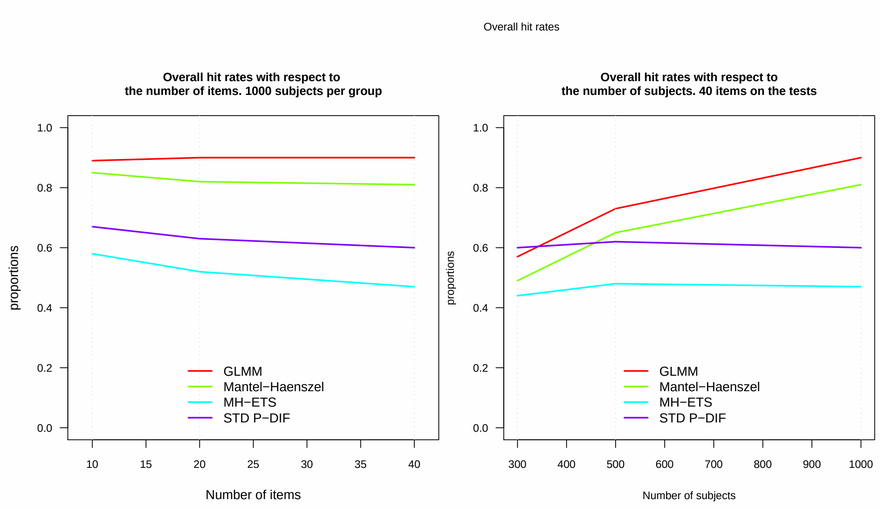
Results and conclusions

NOT ALL SETTINGS ARE SHOWN IN THIS POSTER

GLMM performance. For all settings:
- **HR are higher**
- **FP stay still under 0.06**
- **The trade-off FP vs. FN decrease with the number of subjects**

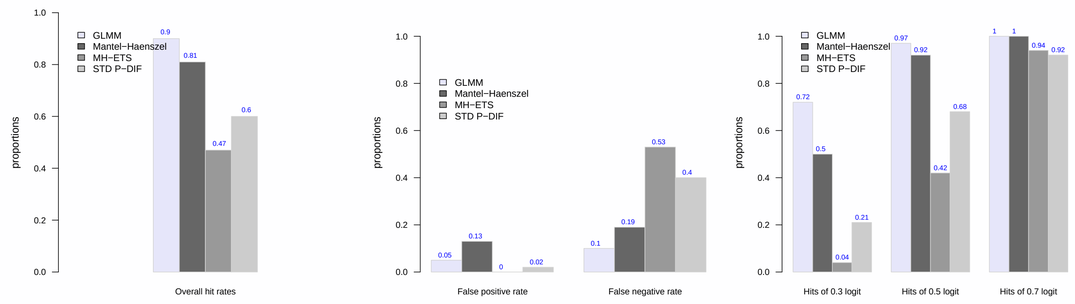






Regarding the evolution with respect to subject or items, GLMM detects higher number of DIF items

For each DIF introduced $\{03, 0.5, 0.7\}$ GLMM HR are higher, in particular for the smallest 0.3 logit



Figures: 1000 subjects per group, 40 items, 20% de DIF items

Advantages and disadvantages of the method

- Flexible: Possible to extend it to other models. Introduction of factors and interactions are straightforward
- No parameter estimation on separate groups needed
- No equating needed
- No anchor needed
- Today, many softwares implement mixed models, in particular R

- GLMM method requires the population to be $\sim \mathcal{N}(0, \sigma^2)$
  → COMPUTATIONAL INTENSIVE
- Mixed models are delicate to manipulate (... but random-intercept model is one of the simplest structures)

Bibliography

The most important, among many others articles
- Differential Item Functioning, P. W. Holland and H. Wainer, Hillsdale, NJ: Lawrence Erlbaum (1993)
- Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach, Eds. Paul De Boeck and Mark Wilson, Series: Statistics for Social and Behavioral Sciences (2004)