Statistics for detecting DIF among multiple groups: A simulation study

Hernández, A.
(University of Valencia, Spain)

Fidalgo, A.M.
(University of Oviedo, Spain)

12th European Congress of Psychology. Istanbul, July 4-8, 2011

Partially supported by the Spanish Ministry of Education and Science CONSOLIDER-C (SEJ2006-14086/PSIC), and by the Spanish Ministry of Science and Innovation (Research project PSI2009-08529).

Overview

- DIF is a potential threat to comparing scores of people belonging to different groups
- Many statistics and procedures for testing DIF when there are two comparison groups (RG and FC)
- But in some cases we need to compare multiple groups
  - cross cultural research
  - multilingual research
  - interactions between two relevant grouping variables

- Generalized Mantel-Haenszel (GMH)
- CFA with latent Mean & Covariance Structure (MACS)

Objective

- Compare the adequacy of GMH and MACS to test DIF in polytomous items across multiple groups:
  - Can be more adequate for relatively small sample sizes than some othe procedures based on IRT
  - Global comparison can be made, no need to compare groups two by two

- Montecarlo simulation to test power and type I error rates of both procedures

Multiple group GMH

- GMH across multiple groups (Q:R*2) (Penfield, 2001)
  - Drawback: Limited to dichotomous items

- Recent extension for polytomous items (Q:R*C) (Fidalgo & Madeira, 2008)

$$Q_{GMH} = \left\{ \sum_{h=1}^{Q} (\mathbf{n}_h - \mathbf{m}_h)' \mathbf{A}_h' \right\} \left\{ \sum_{h=1}^{Q} \mathbf{A}_h \mathbf{V}_h \mathbf{A}_h' \right\}^{-1} \left\{ \sum_{h=1}^{Q} \mathbf{A}_h (\mathbf{n}_h - \mathbf{m}_h) \right\}$$

$$where\ \mathbf{A}_h = \mathbf{C}_h \otimes \mathbf{R}_h$$

	Response Variable Categories						
Factor levels	1	2	.	j	.	C	Total
1	n_{h11}	n_{h12}	.	n_{h1j}	.	n_{h1C}	$N_{h1.}$
2	n_{h21}	n_{h22}	.	n_{h2j}	.	n_{h2C}	$N_{h2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	n_{hi1}	n_{hi2}	.	n_{hij}	.	n_{hiC}	$N_{hi.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
R	n_{hR1}	n_{hR2}	.	n_{hRj}	.	n_{hRC}	$N_{hR.}$
Total	Nh·1	Nh·2		Nh·j		Nh·C	Nh

- MG-GMH Simulation studies (Fidalgo & Scalon, 2010) : MG-GMH preferable to multiple pair-wise tests (even after Bonferroni)
  - Well-controlled Type I error
  - Equal or better power, especially for uniform DIF

Multiple group CFA-MACS

$$x_{ij}^{(g)} = \mu_j^{(g)} + \lambda_j^{(g)} \xi_i^{(g)} + \delta_{ij}^{(g)}$$

Location — Discrimination

- An item is DIF free if both parameters are invariant
- Hypothesis of invariance is typically tested by comparing significance of χ^2 for nested models

Simulation studies

(Stark et al., 2006, González-Romá et al., 2006; Hernández et al., 2008, Meade & Lautenshlager, 2004)

- Power generally adequate, especially for uniform DIF
- Type I error depends on the baseline model taken for comparison: Fully free or Fully constrained (better in 1st case)

Objective

- Montecarlo simulation to test power and type I error rates in detecting DIF in polytomous graded items of Multiple group GMH and MACS, when there are more than two groups.

  - MG-GMH

  - MG-MACS: Most efficient version that starts with the fully constrained baseline model and uses the MIs to flag DIF items

- Two possibilities: Applying Bonferroni correction for the number of items evaluated or not

Simulation conditions

- Took the parameters used by González-Romá et al. (2006) to simulate the data (used MACS model: generated continuous data and categorize afterwards)

  - 3 groups, equal latent distributions
  - 10 items with 5 graded response categories
  - One DIF item in one group
  - Four DIF conditions
    - None, small, medium, large (differences in item intercepts equal to 0, .10, .25 and .50)
  - Two sample size conditions
    - 100/100 and 400/400
  - 100 replications

Results

- MACS showed too high type I error rates if no Bonferroni correction was applied (25%)

- GMH showed too low power if Bonferroni correction was applied

↓

Comparison of the best results of both procedures

Results

			Power	Type I
400	MACS	None	---	0,024
		Small	0,270	0,032
		Medium	1,000	0,043
		Large	1,000	0,087
	GMH	None	---	0,048
		Small	0,310	0,053
		Medium	1,000	0,051
		Large	1,000	0,042

Results

			Power	Type I
100	MACS	None	---	0,035
		Small	0,050	0,029
		Medium	0,500	0,024
		Large	1,000	0,029
	Generalized MH	None	---	0,049
		Small	0,060	0,043
		Medium	0,430	0,037
		Large	0,980	0,041

Conclusions

- When applying MACS taking the fully-constrained baseline model: Bonferroni correction for the number of items analyzed is recommended
- When applying GMH Bonferroni correction should not be applied
- If recommendations are followed, both MACS and GMH
  - Good control for the type I error (MACS slightly worse with large DIF and bigger sample sizes)
  - Very high power with small samples if DIF is large
  - Very high power when DIF is medium if sample sizes are moderate

Conclusions

- LIMITATIONS AND FUTURE RESEARCH:
  - The conditions are limited: New studies under extended conditions
  - The model used to generate the data was MACS, which could affect the results
  - Despite the limitations results suggest that both procedures are efficient approaches to test DIF (at least uniform DIF) across more than two groups

Statistics for detecting DIF among multiple groups: A simulation study

Hernández, A.
(University of Valencia, Spain)

Fidalgo, A.M.
(University of Oviedo, Spain)

12th European Congress of Psychology. Istanbul, July 4-8, 2011

Partially supported by the Spanish Ministry of Education and Science CONSOLIDER-C (SEJ2006-14086/PSIC), and by the Spanish Ministry of Science and Innovation (Research project PSI2009-08529).