Differential Item Functioning in a Large-Scale Preschool Science Assessment
Ariela C. Greenberg, Randall D. Penfield, Daryl B. Greenfield

Introduction

- Early science education can provide a strong foundation for school readiness and future learning.
- When assessing early science skills, it is important that the scores obtained from such assessments are valid.
- One way to address validity is by examining the internal structure of items, through analyses such as differential item functioning (DIF).
- Two widely used approaches to estimate DIF in dichotomously scored items are logistic regression (LR) and the Mantel-Haenszel log-odds ratio (MH-LOR).
- In theory, where nonuniform DIF is not present, LR and MH-LOR estimates for uniform DIF effect size should be equivalent.
- This equivalence has been studied in simulated data, but much less in applied settings.

Objective

- This study aimed to provide evidence of validity - based on internal structure by using DIF analysis - for scores obtained from the Nature and Science scales of the Galileo System for the Electronic Management of Learning (Galileo).
- This study also explored the extent to which LR and MH-LOR provided similar estimates of uniform DIF in an applied setting.

Studied Measure

- Florida Head Start teachers use Galileo to track student mastery of specific skills and school readiness across several domains including science.
- The Nature and Science scale has 57 dichotomous items.
- The total score correct is used as the matching criteria in the DIF procedures.

Participants

- 3,636 children ages 4-5 years old from Florida Head Start programs
  - 1875 males (51.6%)
  - 1761 females (48.4%)
  - 479 European-American (13.2%)
  - 1760 African-American (48.4%)
  - 1176 Hispanic (32.3%)

Procedure

- Three DIF analyses were conducted comparing the following groups:

1) European-American / African-American
2) European-American / Hispanic-American
3) Male / Female

- **Logistic Regression** (LR) models were tested to detect nonuniform and uniform DIF in all 57 items.

Model 3 $\quad \psi = \delta + \tau_1 X + \tau_2 G + \tau_3 XG$

Model 2 $\quad \psi = \delta + \tau_1 X + \tau_2 G$

Model 1 $\quad \psi = \delta + \tau_1 X$

where τ_2 = uniform DIF effect; δ = intercept; ψ = LOR; X = observed score; G = group

- **Mantel-Haenszel log-odds ratio** (MH-LOR) method was used to detect uniform DIF in all 57 items.

- LR and MH-LOR estimates of uniform DIF were compared separately for items
  - without nonuniform DIF
  - with nonuniform DIF

- Three types of comparisons:
  - Pearson correlations between estimates
  - Effect size estimates and percentage of congruent results
  - Mean difference and mean absolute difference between estimates

Results

DIF Items - "C" in at least one method

DIF analysis	Item	Description	τ_2	SE	χ^2	ETS (ESO)	β_2	SE	ETS
			Logistic Regression Model 2 uniform DIF coefficient				MH-LOR uniform DIF estimate		
HA (-)	17	Distinguishes plants from animals.	-0.68	0.20	10.96**	C	-0.71	0.21	B
AA (-)	23	Classifies objects by their state (e.g., liquid, solid, gas).	-0.69	0.13	30.60**	C	-0.66	0.13	B
AA (-)	33	Predicts that plants may be destroyed when an animal population increases.	-0.86	0.24	13.90*	B [C]	-0.77	0.27	B
HA (-)	33	Predicts that plants may be destroyed when an animal population increases.	-0.69	0.25	8.45**	C	-0.74	0.29	B
AA (+)	38	Predicts that a rolling object will move faster than a sliding object.	0.65	0.13	24.44**	B [C]	0.69	0.13	C
AA (-)	47	Presents observations in a variety of ways (e.g., drawings, charts, maps).	-0.68	0.16	18.63**	B [C]	-0.71	0.18	B
AA (-)	48	Describes her/his data to other peers/adults.	-0.73	0.15	25.85**	C	-0.76	0.16	C

Note: ETS, using full Educational Testing Service Criteria: A <.43 or not significantly different from 0; .43 < B < .64, significantly different from 0, but not .43; C > .64 and significantly different from .43); ESO, using effect size only to categorize

Example of comparing LR and MH-LOR approaches: European-American / African-American analysis

Correlations of LR coefficients with MH-LOR

$r = .99****$
$r = .99***$
$r = .395$
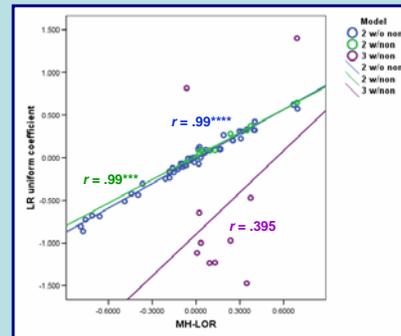
Mean differences and absolute mean differences

	Difference between MH-LOR and LR uniform coefficient in	N items	Mean difference	SD	Absolute Mean difference
items without nonuniform DIF	Model 2	47	-0.011	0.049	0.04
items with nonuniform DIF	Model 2	10	0.002	0.035	0.027
	Model 3	10	-0.779	0.886	1.097

Congruency patterns and rates

Items without nonuniform DIF — LR Model 2

MH-LOR	A	B	C
A	38	0	0
B	1	5[3]	2[4]
C	0	0	1

94 [90]% congruent

Items with nonuniform DIF — LR Model 2

MH-LOR	A	B	C
A	9	0	0
B	0	0	0
C	0	1[0]	0[1]

90 [100]% congruent

Items with nonuniform DIF — LR Model 3

MH-LOR	A	B	C
A	0	9[1]	0[8]
B	0	0	0
C	0	0	1

10% congruent

[ ] using effect size only

Conclusions

- In the gender comparison, DIF was not detected.
- In the ethnicity studies, overall, most items do not display meaningful DIF.
- In the ethnicity studies, 7 items showed meaningful DIF, six of which favored the focal group, and will be reviewed for content by experts in early science education for possible item bias, which could impact validity of the assessment scores.

- For items without nonuniform DIF, the LR Model 2 uniform DIF coefficient and MH-LOR were highly congruent, which was consistent with the literature on simulated studies.
- Even for items with nonuniform DIF, the LR Model 2 uniform DIF coefficient was more congruent with MH-LOR than the LR Model 3 uniform coefficient.
- When only interested in uniform DIF effects of items, the MH-LOR is a more efficient method than LR.