Lesson 5: Differential Item Functioning

Patrícia Martinková

Department of Statistical Modelling
Institute of Computer Science, Czech Academy of Sciences

NMST 570, December 12, 2017
Contents

1. Introduction

2. DIF and fairness

3. DIF detection methods

4. difNLR

5. ShinyItemAnalysis

6. Conclusion
Motivation: Development and Validation of HCl

Complex validation of Homeostasis Concept Inventory (HCl)

---

Motivation: Development and Validation of HCl

Complex validation of Homeostasis Concept Inventory (HCI)

- Males / English as a first language / White and Asian students performed better

---

doi 10.1187/cbe.16-10-0305
Motivation: Development and Validation of HCl

Complex validation of Homeostasis Concept Inventory (HCI)

- Males / English as a first language / White and Asian students performed better

Is the test fair?

__________________________________________________________________________________

doi 10.1187/cbe.16-10-0305
Motivation: Development and Validation of HCI

Differential Item Functioning (DIF) Analysis

- Analytical method to address item fairness
- Ubiquitous in large-scale assessments development
- Less used in conceptual assessment development

Martinková et al. Checking Equity: Why DIF Analysis should be a Routine Part of Developing Conceptual Assessments. *CBE Life Sciences Education*, 16(2), rm2. doi 10.1187/cbe.16-10-0307
Motivation: Development and Validation of HCI

Differential Item Functioning (DIF) Analysis

- Analytical method to address item fairness
- Ubiquitous in large-scale assessments development
- Less used in conceptual assessment development

- None of the HCI items exhibited DIF
  - with respect to gender, ethnicity or ELL status

Martinková et al. Checking Equity: Why DIF Analysis should be a Routine Part of Developing Conceptual Assessments. *CBE Life Sciences Education*, 16(2), rm2. doi 10.1187/cbe.16-10-0307
Motivation: Development and Validation of HCI

Differential Item Functioning (DIF) Analysis

- Analytical method to address item fairness
- Ubiquitous in large-scale assessments development
- Less used in conceptual assessment development

None of the HCI items exhibited DIF

- with respect to gender, ethnicity or ELL status

Methods paper: Importance of DIF Analysis

Martinková et al. Checking Equity: Why DIF Analysis should be a Routine Part of Developing Conceptual Assessments. *CBE Life Sciences Education*, 16(2), rm2.
doi 10.1187/cbe.16-10-0307
Differential Item Functioning (DIF)

Two subjects with the same underlying ability but from different groups have different probability to answer question correctly.
Differential Item Functioning (DIF)
Two subjects with the same underlying ability but from different groups have different probability to answer question correctly

- Two groups referred to as reference and focal (usually minority)
Differential Item Functioning (DIF)
Two subjects with the same underlying ability but from different groups have different probability to answer question correctly

- Two groups referred to as reference and focal (usually minority)
- Two types of DIF - uniform and non-uniform
Example of DIF item

Childhood illnesses (Drabinová & Martinková, 2017)

Deficiency of vitamin D in childhood could cause

VITAMIN D
MORE THAN JUST STRONG BONES
Example of DIF item

Childhood illnesses (Drabinová & Martinková, 2017)

Deficiency of vitamin D in childhood could cause

a. rickets
Example of DIF item

Childhood illnesses (Drabinová & Martinková, 2017)

Deficiency of vitamin D in childhood could cause

- rickets
- scurvy
Example of DIF item

Childhood illnesses (Drabinová & Martinková, 2017)

Deficiency of vitamin D in child-
hood could cause

a. rickets
b. scurvy
c. dwarfism
Example of DIF item

Childhood illnesses (Drabinová & Martinková, 2017)

Deficiency of vitamin D in childhood could cause

a. rickets
b. scurvy
c. dwarfism
d. mental retardation
Example of DIF item

Tipping example (Martiniello et al., 2012)

Of the following, which is the closest approximation of a 15 percent tip on a restaurant check of $24.99?

a. $2.50
b. $3.00
c. $3.75
d. $4.50
Example of DIF items

- Example: Spelling test (orally administered)
  - spell word *girder*
Example of DIF items

- Example: Spelling test (orally administered)
  - spell word *girder*

- Example (SAT): Runner is to marathon as
  - a. envoy is to embassy
  - b. martyr is to massacre
  - c. oarsman is to regatta
  - d. referee is to tournament
  - e. horse is to stable
Example of DIF items

- Example: Spelling test (orally administered)
  - spell word girder

- Example (SAT): Runner is to marathon as
  a. envoy is to embassy
  b. martyr is to massacre
  c. oarsman is to regatta
  d. referee is to tournament
  e. horse is to stable
Example of DIF items

- Example: Spelling test (orally administered)
  - spell word girder

- Example (SAT): Runner is to marathon as
  a. envoy is to embassy
  b. martyr is to massacre
  c. oarsman is to regatta
  d. referee is to tournament
  e. horse is to stable

Who might have been disadvantaged?
Terminology: Reference group (R), Focal group (F)
DIF as multidimensionality problem:

- Existence of another dimension tested on the particular item besides the primary latent variable
DIF as multidimensionality problem:

- Existence of another dimension tested on the particular item besides the primary latent variable

What is the primary and the secondary latent variable tested in mentioned examples?
DIF and item fairness

DIF items are potentially unfair

- Content experts must decide on item fairness
DIF and item fairness

DIF items are **potentially** unfair

- Content experts must decide on item fairness
- Secondary latent trait causing DIF
DIF and item fairness

DIF items are **potentially** unfair

- Content experts must decide on item fairness
- Secondary latent trait causing DIF
  - Unrelated to content being tested
    - DIF item is considered unfair
    - Item should be reworded or removed
    - Example: Tipping
DIF and item fairness

DIF items are **potentially** unfair

- Content experts must decide on item fairness
- Secondary latent trait causing DIF
  - Unrelated to content being tested
    - DIF item is considered unfair
    - Item should be reworded or removed
    - Example: Tipping
  - Related to content being tested
    - DIF item is not considered unfair
    - Item can inform teaching
    - Example: Item on childhood illnesses as part of Czech Medical School Admission Test in Biology
Comparing total scores only can lead to incorrect conclusions about item/test fairness:

- **Case study 1: Homeostasis Concept Inventory**
  - Significant difference between males and females in total score (Fig A)
- **Case study 2: Simulated dataset based on GMAT**
  - Identical distributions of total score (Fig B)

---

Martinková et al. (2017)
DIF vs. Difference in total scores (cont.)

Comparing total scores only can lead to incorrect conclusions about item/test fairness:

- Case study 1: No HCl item detected as DIF

Martinková et al. (2017)
Comparing total scores only can lead to incorrect conclusions about item/test fairness:

- Case study 1: No HCI item detected as DIF
- Case study 2: DIF detected in two items of simulated dataset
  - Item 1 exhibits uniform DIF (Fig A)
  - Item 2 exhibits non-uniform DIF (Fig B)

Martinková et al. (2017)
### DIF detection methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mantel-Haenszel test</td>
<td>Simple, easily implemented, detects both forms of DIF</td>
<td>Cannot detect non-uniform DIF, doesn't account for possibility of guessing/inattention</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>Simple, easily implemented, detects both forms of DIF</td>
<td>Doesn't account for possibility of guessing/inattention</td>
</tr>
<tr>
<td>Item Response Theory models (non-linear mixed effects)</td>
<td>Detects both forms of DIF, accounts for possibility of guessing/inattention</td>
<td>More complex, computationally demanding</td>
</tr>
</tbody>
</table>

Patrícia Martinková

NMST570, L5: Differential Item Functioning

December 12, 12/26
DIF detection methods

- Based on total score
DIF detection methods

- **Based on total score**

- **Based on latent ability**
DIF detection methods

- Based on total score
  - Mantel-Haenszel test
    - simple, easily implemented
    - cannot detect non-uniform DIF
    - doesn’t account for possibility of guessing/inattention

- Based on latent ability
DIF detection methods

- Based on **total score**
  - Mantel-Haenszel test
    - simple, easily implemented
    - cannot detect non-uniform DIF
    - doesn’t account for possibility of guessing/inattention
  - Logistic regression
    - simple, easily implemented, detects both forms of DIF
    - doesn’t account for possibility of guessing/inattention

- Based on **latent ability**
DIF detection methods

• Based on **total score**
  • Mantel-Haenszel test
    • simple, easily implemented
    • cannot detect non-uniform DIF
    • doesn’t account for possibility of guessing/inattention
  • Logistic regression
    • simple, easily implemented, detects both forms of DIF
    • doesn’t account for possibility of guessing/inattention

• Based on **latent ability**
  • Item Response Theory models (non-linear mixed effect models)
    • detects both forms of DIF, accounts for possibility of guessing/inattention
    • more complex, computationally demanding
Mantel-Haenszel test

- Test of independence of two binary variables: item score and group membership.
- $X^2$ test, but incorporating also ability score
- Looking at contingency tables for each level of total score, adding up
Logistic regression for DIF detection

\[ P(Y_{ij} = 1 | X_i, G_i) = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}} \]

= probability of correct answer of student \( i \) to item \( j \)

\( X_i \) total score, \( G_i \) group
Logistic regression for DIF detection

\[
P(Y_{ij} = 1|X_i, G_i) = \frac{e^{\beta_0 j + \beta_1 j X_i + \beta_2 j G_i}}{1 + e^{\beta_0 j + \beta_1 j X_i + \beta_2 j G_i}}
\]

= probability of correct answer of student \(i\) to item \(j\)

\(X_i\) total score, \(G_i\) group
Logistic regression for DIF detection

\[
P(Y_{ij} = 1|X_i, G_i) = \frac{e^{\beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_3 X_i G_i}}{1 + e^{\beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_3 X_i G_i}}
\]

= probability of correct answer of student \(i\) to item \(j\)

\(X_i\) total score, \(G_i\) group

![Graph showing the probability of correct answer against total score for two groups.](image-url)
IRT-based Methods for DIF Detection

- Lord’s Wald statistic: Difference between parameters
- Raju: Area between the curves (difference or absolute difference)
- Likelihood ratio test

![Difference of estimated parameters](image1)

![Area between curves](image2)
Generalized logistic regression for DIF detection

\[ P(Y_{ij} = 1 \mid X_i, G_i) = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}} \]

= probability of correct answer by \(i\)th subject on \(j\)th item

\(X_i\) total score, \(G_i\) group membership

\[ \text{Drabinová & Martinková (2017)} \]
Generalized logistic regression for DIF detection

\[ P(Y_{ij} = 1 | X_i, G_i) = c_j + (d_j - c_j) \frac{e^{\beta_0j + \beta_1jX_i}}{1 + e^{\beta_0j + \beta_1jX_i}} \]

= probability of correct answer by \( i \)th subject on \( j \)th item

\( X_i \) total score, \( G_i \) group membership

---

Drabinová & Martinková (2017)
Introduction

DIF and fairness

DIF detection methods

difNLR

ShinyItemAnalysis

Conclusion

Generalized logistic regression for DIF detection

\[ P(Y_{ij} = 1 | X_i, G_i) = c_j + (d_j - c_j) \frac{e^{\beta_0 j + \beta_1 j X_i + \beta_2 j G_i + \beta_3 j X_i G_i}}{1 + e^{\beta_0 j + \beta_1 j X_i + \beta_2 j G_i + \beta_3 j X_i G_i}} \]

= probability of correct answer by \( i \)th subject on \( j \)th item

\( X_i \) total score, \( G_i \) group membership

Drabinová & Martinková (2017)
Technical details

We use:

- Z-scores instead of total score
- IRT parameterization
- Non-linear least squares for parameter estimation
- DIF testing based on F or LR test
- Multiple comparison corrections

https://CRAN.R-project.org/package=difNLR
We use:

- Z-scores instead of total score
- IRT parameterization
- Non-linear least squares for parameter estimation
- DIF testing based on F or LR test
- Multiple comparison corrections

Method is implemented in R library difNLR (Drabinová, Martinková & Zvára, 2017)

https://CRAN.R-project.org/package=difNLR
Different asymptotes for groups

- Model allows for differences in guessing between groups

Monte Carlo simulation study

Design

- 5 levels of sample size
  (500+500, 500+1,000, 1,000+1,000, 1,000+2,000, 2,000+2,000)
- 20 items
- Answers generated using 3PL model
- DIF caused by difference in difficulty, discrimination and guessing parameters
- 0%, 5%, or 15% DIF proportion
- DIF size based on (weighted) area between characteristic curves
Monte Carlo simulation study

Design
- 5 levels of sample size
  (500+500, 500+1,000, 1,000+1,000, 1,000+2,000, 2,000+2,000)
- 20 items
- Answers generated using 3PL model
- DIF caused by difference in difficulty, discrimination and guessing parameters
- 0%, 5%, or 15% DIF proportion
- DIF size based on (weighted) area between characteristic curves

DIF detection
- Mantel-Haenszel, Logistic Regression, Lord (3PL IRT), NLR
- Benjamini-Hochberg multiple comparison correction
Monte Carlo simulation study

Results - NLR

- Less convergence issues than for Lord (3PL IRT)
- Good control of rejection rates in almost all scenarios
- Comparable power to other DIF detection methods
- Accounts for guessing
- Allows for testing group difference in guessing

Differential Distractor Functioning (DDF)

Two subjects with the same underlying ability but from different groups have different probability to choose given distractor in multiple-choice item.

Martinková & Drabinová, in progress.
DDF for detection of differential attractiveness of distractors

Extending multinomial regression model
- To better describe attractiveness of distractors

Extending DDF model
- To account for differential attractiveness of distractors in multiple-choice items

Martinková & Drabinová, in progress.
Simulated GMAT data: total scores may have exactly the same distribution, yet there may be DIF present in some items!

Martinková et al. Checking Equity: Why DIF Analysis should be a Routine Part of Developing Conceptual Assessments. *CBE Life Sciences Education*, 16(2), rm2. doi 10.1187/cbe.16-10-0307
Method demonstrated on MSAT-B dataset from Drabinová & Martinková (2017)

ShinyItemAnalysis: DDF with multinomial regression
Conclusion

DIF/DDF analysis should be used routinely in test development

- to check for fairness with respect to groups
- to inform teaching
Conclusion

DIF/DDF analysis should be used routinely in test development

- to check for fairness with respect to groups
- to inform teaching

DIF detection methods

- Mantel-Haenszel test
- Logistic regression
- IRT/based methods: Lord (Wald test), Raju
Conclusion

DIF/DDF analysis should be used routinely in test development

- to check for fairness with respect to groups
- to inform teaching

DIF detection methods

- Mantel-Haenszel test
- Logistic regression
- IRT/based methods: Lord (Wald test), Raju

New method for DIF detection was introduced

- allows for group differences in guessing and inattention
- current research focuses on differences in option selection (DDF)
- may provide better understanding to misconceptions held by groups
Thank you for your attention!

www.cs.cas.cz/martinkova
McFarland, Price, Wenderoth, Martinková, Cliff, Michael, Modell and Wright (2017). Development and Validation of the Homeostasis Concept Inventory. *CBE Life Sciences Education*, vol. 16 no. 2 ar35. doi 10.1187/cbe.16-10-0305


