Assessment, analysis and interpretation of Patient-Reported Outcomes (PROs)

Day 5

Summer school in Applied Psychometrics

Peterhouse College, Cambridge

12th to 16th September 2011

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

This course is prepared by

Anna Brown, PhD ab936@medschl.cam.ac.uk

Jan Stochl, PhD js883@cam.ac.uk

Tim Croudace, PhD tjc39@cam.ac.uk

(University of Cambridge, department of Psychiatry)

Jan Boehnke, PhD boehnke@uni-trier.de

(University of Trier, Department of Clinical Psychology and Psychotherapy)

The course is funded by the ESRC RDI and hosted by The Psychometrics Centre

Jan Boehnke

13. MEASUREMENT INVARIANCE

Agenda

- Measurement equivalence and types of bias
- Concepts of item impact, Differential Item Functioning (DIF) and item bias
- Investigating DIF using
  - non-parametric techniques (Mantel-Haenszel),
  - parametric techniques (logistic regression).

Levels of measurement equivalence

- Structural / functional equivalence
  - The same psychological constructs is measured across groups (for example, patterns of correlations between variables are the same across groups)
- Measurement unit equivalence
  - The same measurement unit (individual differences found in group A can be compared with differences found in group B)
- Scalar / full score equivalence
  - The same measurement unit and the same origin (scores can be compared across groups)

Van de Vijver & Poortinga

Types of bias

- Construct bias
  - Definition/appropriateness of constructs is different between cultures
- Method bias
  - Instrument bias – instrument features not related to the construct (familiarity with stimulus material etc.)
  - Administration bias
  - Response bias
- Item bias
  - Poor translation
  - Item-related nuisance factors (e.g. item may invoke additional traits or abilities)
- Sample bias
  - demographics mix - balance of demographics within samples may differ

Influence of bias on the level of equivalence

Type of Bias	Structural equivalence	Measurement unit equivalence	Scalar equivalence
Construct bias	yes	yes	yes
Method bias: uniform	no	no	yes
Method bias: non-uniform	no	yes	yes
Item bias: uniform	no	no	yes
Item bias: non-uniform	no	yes	yes

Item impact and DIF

- Item impact is evident when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item
  - Can be because there are true differences between the groups in the underlying construct
  - Or because the item is biased (unfair to one group)
- Differential Item Functioning (DIF) occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item *after matching on the underlying construct* that the item is intended to measure

Item bias

- Item bias occurs when examinees of one group are less likely to answer an item correctly (or endorse an item) than examinees of another group because of some characteristic of the test item that is *not relevant* to the construct being measured

Item impact & bias

- Analyses of item bias are *qualitative*: reconstruction of meaning and contextualization
- Analyses of DIF are usually statistical in nature: testing whether differences in probabilities remain, when matched on trait level

- DIF is required, but not sufficient, for item bias.
  - If no DIF is apparent, there is no item bias
  - If DIF is apparent, additional investigations are necessary (e.g. content analysis by subject matter experts)

Item bias or item impact?

- Example 1. Students are asked to compare the weights of several objects, including a football.
  - Since girls are less likely to have handled a football, they found the item more difficult than boys, even though they have mastered the concept measured by the item (Scheuneman, 1982a).
- Example 2. A vocabulary test asked to find a synonym to "ebony".
  - The Black students were more likely to answer the item correctly than the White students throughout the bulk of the test score distribution. Ebony is a dark-coloured wood and it is also the name of a popular magazine targeted to African-Americans.
  - The item was considered to an important part of the curriculum and was not removed from the test.

DIFFERENTIAL ITEM FUNCTIONING

"Sample Free"

Fayers & Machin (2007), p. 164:

"Another important aspect of IRT is that it is 'sample free', because the *relative* item difficulties should remain the same irrespective of the particular sample of subjects." (italics by authors)

What does that mean?

"Sample Free"

- The notion of sample free estimates has been misunderstood often
- sample free says that: "when an IRT model holds in a population, then any sample from a subgroup of this population should lead to the same estimates of the item parameters"
- whether this is true (within range of sampling error) is an empirical question!

Purposes of DIF studies

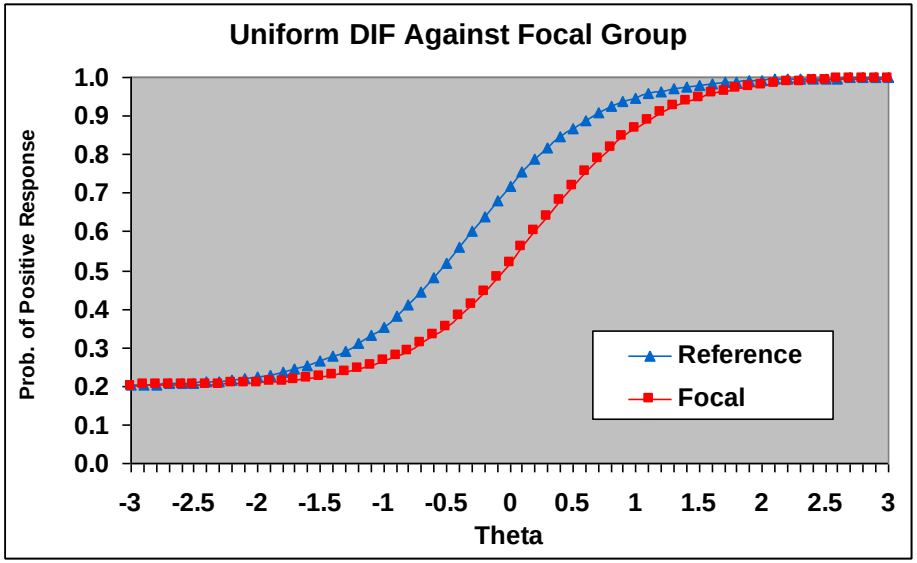
- *Purpose 1: Fairness and equity in testing.*
- *Purpose 2: Dealing with a possible threat to internal validity.*
  - rule out measurement artifact as an explanation for the group differences
- *Purpose 3: Investigate the comparability of translated and/or adapted measures.*
- *Purpose 4: Trying to understand item response processes.*
- *Purpose 5: Investigating lack of invariance.*

Zumbo, B. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going.

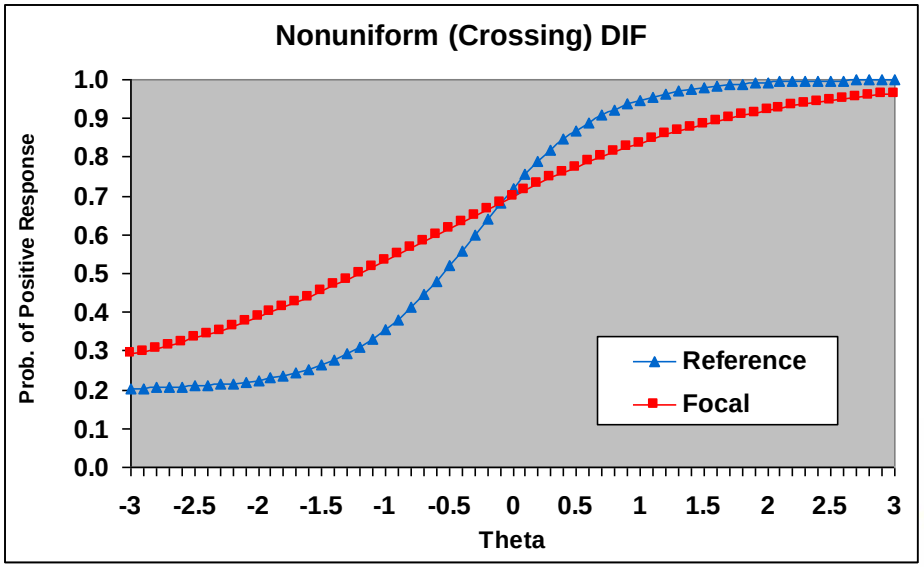
Terminology

- Reference and focal groups
  - The reference group is the group that serves as the standard
  - The focal group is the group that is compared against the standard
  - Typically, the majority group or the group on which a test was standardized serves as the reference group
- Matching variable
  - Participants from the different groups are matched with respect to their proficiency. The matching variable is the variable that represents the latent construct
  - It can be operationalized as the total test score, or IRT estimated ability (depending on method)

Uniform and non-uniform DIF



Focal group has lower probability of endorsing the item at all trait levels



Focal group has higher probability of endorsing the item at low level of trait, but lower probability at high level

Differential Test Functioning

- Differential test functioning (DTF) is present when individuals who have the same standing on the latent construct or attribute, but belong to different groups, obtain different scores on the test
- The presence of DIF may lead to DTF, but not always
  - some DIF items favour the focal group, whereas others may favour the reference group, which produces a cancelling effect
- DTF is of greater practical significance than DIF
- Ideally, we want a test with no DIF and no DTF

Types of DIF techniques

- Non-parametric
  - Mantel-Haenszel statistic and its variations (Holland & Thayer, 1988)
  - TestGraf (non-parametric IRT; Ramsay 1994)
  - Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993)
- Parametric
  - Logistic regression (Swaminathan & Rogers, 1990)
  - Item Response Theory methods
  - Structural Equation Modelling (e.g. Muthen & Lehman, 1985)

Three pieces of information necessary for DIF analysis

- Group membership
- Score on a matching variable
- Response to an item
  - DIF is present when expected item scores differ across groups conditional on the matching variable
  - DIF is present when group membership tells one something about responses to an item after controlling for the latent construct

Non-parametric DIF technique

BINARY MANTEL-HAENSZEL

The Mantel-Haenszel method

- A popular DIF method since the late 1980's; still stands as very effective compared with newer methods
- Used by Educational Testing Service (ETS) in screening for DIF
- The MH method treats the DIF detection problem as one involving three-way contingency tables. The three dimensions of the contingency table involve
  - (a) whether one gets an item correct or incorrect
  - (b) group membership, while conditioning on
  - (c) the total score "sliced" into a number of category score bins.

Score “slices”

- The total score is divided into score groups (slices)
- Slices may be “thin” or “thick” depending on the sample size
- With many participants the total score can be divided into thin slices
  - Ideally each slice should correspond to a score on the total score scale
  - For instance, if the total score ranges from 0 to 10, there will be eleven score groups

Chi-square contingency table

Performance on an item *at score level (slice) j*

	1	0	
Reference group	a_j	b_j	$N_{Rj} = a_j + b_j$
Focal group	c_j	d_j	$N_{Fj} = c_j + d_j$
	$N_{1j} = a_j + c_j$	$N_{0j} = b_j + d_j$	$N_j = a_j + b_j + c_j + d_j$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Mantel-Haenszel statistic

$$MH = \frac{\left(\left| \sum_j a_j - \sum_j E\left(a_j\right) \right| - 0.5 \right)^2}{\sum_j \mathrm{var}\left(a_j\right)}$$

- Where $E\left(a_j\right) = \frac{N_{Rj} N_{1j}}{N_j}$ $\mathrm{var}\left(a_j\right) = \frac{N_{Rj} N_{1j} N_{Fj} N_{0j}}{N_j^2 \left(N_j - 1\right)}$
- Restricted to the sum over slices that are actually observed in the dataset
- Null hypothesis = no association between item response and group membership
- MH follows a chi-square distribution with 1 degree of freedom and is used for significance testing

Mantel-Haenszel common odds ratio for an item at score level *j*

$$\alpha_j = \frac{p_{Rj}}{q_{Rj}} \Big/ \frac{p_{Fj}}{q_{Fj}} = \frac{a_j d_j}{b_j c_j}$$

Where

p_{Rj} = number of persons in Reference group in score interval j who answered correctly;

q_{Rj} = number of persons in Reference group in score interval j who answered incorrectly.

Notation *F* relates to the focal group

If the item does not show DIF, we expect this ratio to be 1

Mantel-Haenszel common odds ratio for item *i*

- For the slice j $$\alpha_j = \frac{a_j d_j}{b_j c_j}$$
- Across all slices $$\hat{\alpha}_{MH} = \frac{\sum_j a_j d_j / N_j}{\sum_j b_j c_j / N_j}$$
- The logarithm of common odds ratio is normally distributed and is used as effect size measure

$$\lambda_{MH} = \log\left(\hat{\alpha}_{MH}\right)$$

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Interpreting the results of the MH procedure

- Step 1: Examine whether the Mantel-Haenszel statistic is statistically significant
- Step 2: Examine the size of the common odds ratio (the DIF effect size)
- Step 3: Use the ETS classification scheme to judge the practical significance of the DIF (see Penfield & Algina, 2006, p. 307)
  - LOR > 0.64 Large DIF (ETS Class C)
  - LOR > 0.43 Moderate DIF (ETS Class B)
  - LOR < 0.43 Small DIF (ETS Class A)

Examining Differential Test Functioning

- Does DIF translate into differential test functioning (DTF)?
  - The variance of the MH DIF effects may be taken as an indicator of DTF
  - The bigger the variance, the more the test functions differently for the reference and focal groups
  - Penfield and Algina devised a DIF effect variance statistic, τ^2 (tau squared), which may be used as an indicator of DTF

Examining Differential Test Functioning

- Step 4: Examine the DIF effect variance as a measure of differential test functioning (DTF)
  - Small DIF effect variance, $\tau^2 < 0.07$ (about 10% or fewer of the items have LOR < ±0.43)
  - Medium DIF effect variance, $0.07 < \tau^2 < 0.14$
  - Large DIF effect variance, $\tau^2 > 0.14$ (about 25% or more of the items have LOR > ±0.43)
  - These cut points may be adjusted by individual users depending on their own needs, substantive knowledge, and experience in the particular field of interest

MH METHOD WITH DIFAS

DIFAS package

- *DIFAS*, and its corresponding manual, can be can be downloaded free of charge from a website of *Randall Penfield (University of Miami)*

  http://www.education.miami.edu/facultysites/penfield/index.html

- Many thanks to *Dr Deon de Bruin (University of Johannesburg)* for
  - Showing DIFAS at a workshop at SIOPSA
  - Providing the example data

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Synthetic data generated to demonstrate DIF with dichotomous items

- Synthetic data for a 15-item test with 2000 respondents
  - Respondents come from two groups (1000 per group)
  - Dataset courtesy Deon De Bruin, University of Johannesburg
- The data were generated according to the Rasch model
  - All the items have equal slopes (*discrimination* parameters)
  - For six items the difficulty parameters (*b*) was specified to differ across groups
  - Hence, six items demonstrate uniform DIF, but no items demonstrate non-uniform DIF
  - The ability of the two groups is equal

True item difficulty parameters (DIF items highlighted)

Item	Group		Item	Group	
	Reference	Focal		Reference	Focal
Item 1	-2.5	-2.5	Item 9	0.0	0.0
Item 2	-2.3	-1.8	Item 10	0.4	1.4
Item 3	-2.0	-2.0	Item 11	1.0	1.0
Item 4	-1.7	-2.3	Item 12	1.2	0.9
Item 5	-1.5	-1.4	Item 13	1.3	1.4
Item 6	-1.2	-0.2	Item 14	1.9	1.9
Item 7	-0.7	-0.7	Item 15	1.6	2.5
Item 8	-0.1	-0.1			

Source: De Bruin, D. (2008). What do you mean your test is cross-culturally valid? *Workshop presented at SIOPSA, Pretoria, SA.*

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Descriptive statistics for the scale

Group	Mean	SD	Cronbach alpha KR-20
Group 1 (n = 1000)	8.17	7.77	.70
Group 2 (n = 1000)	7.87	7.42	.68
Total (n = 2000)	8.02	7.61	.69

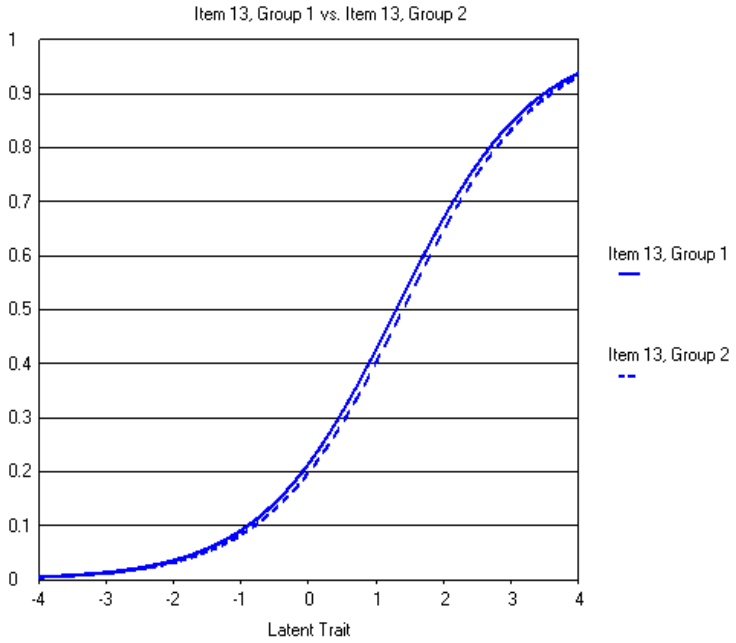
Casual inspection shows similar means, SD's and reliabilities.

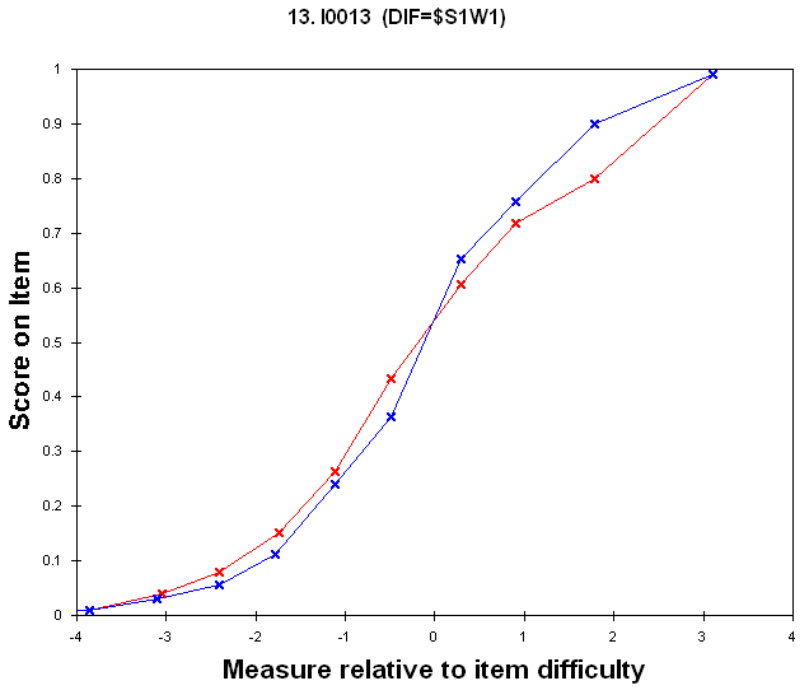
Source: De Bruin, D. (2008). What do you mean your test is cross-culturally valid? *Workshop presented at SIOPSA, Pretoria, SA.*

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Theoretical and empirical IRFs

- Item 13 is designed to show no DIF



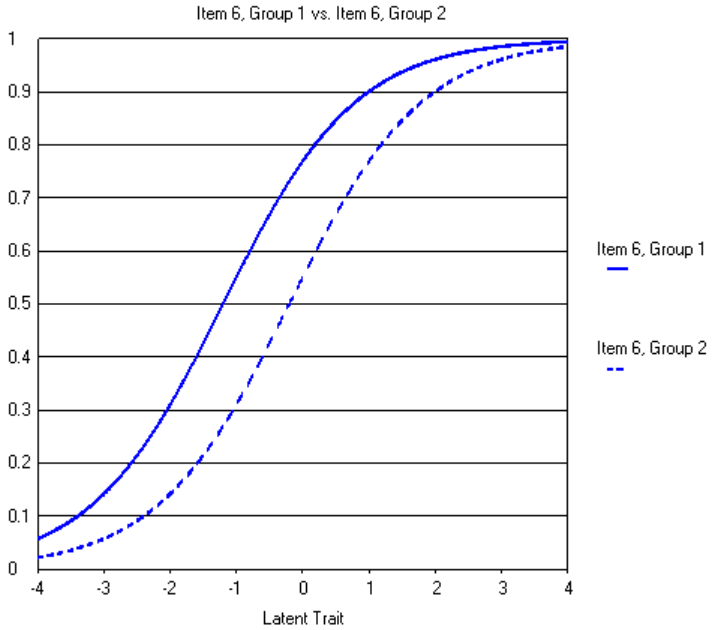


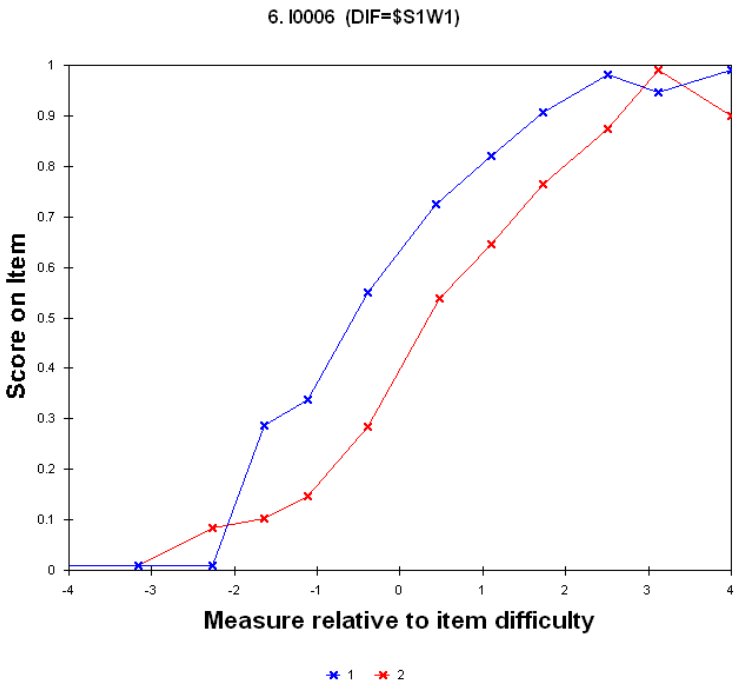
Source: De Bruin, D. (2008). What do you mean your test is cross-culturally valid? *Workshop presented at SIOPSA, Pretoria, SA.*

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Theoretical and empirical IRFs

- Item 6 is designed to show DIF





Source: De Bruin, D. (2008). What do you mean your test is cross-culturally valid? *Workshop presented at SIOPSA, Pretoria, SA.*

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Results of the Mantel-Haenszel test (obtained with DIFAS 5)

DIF STATISTICS: DICHOTOMOUS ITEMS

Name	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
Var 1	0.2461	0.0958	0.1659	0.5775	0.49	OK	A
Var 2	7.658	0.3946	0.1393	2.8327	0.365	Flag	A
Var 3	1.8162	-0.2007	0.1413	-1.4204	0.007	OK	A
Var 4	**32.4658**	**-0.7750**	**0.1374**	**-5.6405**	**0.122**	**Flag**	**C**
Var 5	0.0342	-0.0297	0.1208	-0.2459	0.047	OK	A
Var 6	**82.8232**	**0.9966**	**0.1109**	**8.9865**	**0.47**	**Flag**	**C**
Var 7	0.3814	-0.0713	0.1062	-0.6714	0.484	OK	A
Var 8	0.6644	-0.0898	0.1035	-0.8676	0.393	OK	A
Var 9	4.9067	-0.2356	0.104	-2.2654	0.033	OK	A
Var 10	31.2327	0.6469	0.1151	5.6203	0.204	Flag	B
Var 11	5.8599	-0.2769	0.1119	-2.4745	2.238	Flag	A
Var 12	**33.0494**	**-0.6519**	**0.1137**	**-5.7335**	**6.947**	**Flag**	**C**
Var 13	1.9575	-0.1794	0.1225	-1.4645	0.583	OK	A
Var 14	5.0798	-0.2983	0.1286	-2.3196	0.093	Flag	A
Var 15	**24.6969**	**0.7288**	**0.1458**	**4.9986**	**0.003**	**Flag**	**C**

Source: De Bruin, D. (2008). What do you mean your test is cross-culturally valid? *Workshop presented at SIOPSA, Pretoria, SA.*

Results of the Mantel-Haenszel test (cont.)

DIF STATISTICS: DICHOTOMOUS ITEMS

Name	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
Var 4	**32.4658**	**-0.7750**	**0.1374**	**-5.6405**	**0.122**	**Flag**	**C**
Var 6	**82.8232**	**0.9966**	**0.1109**	**8.9865**	**0.470**	**Flag**	**C**
Var 10	31.2327	0.6469	0.1151	5.6203	0.204	Flag	B
Var 12	**33.0494**	**-0.6519**	**0.1137**	**-5.7335**	**6.947**	**Flag**	**C**
Var 15	**24.6969**	**0.7288**	**0.1458**	**4.9986**	**0.003**	**Flag**	**C**

A negative sign shows item is easier for focal group

LOR > 0.64 moderate to large DIF (ETS C)
LOR > 0.43 slight to moderate DIF (ETS B)
LOR < 0.43 slight DIF (ETS A)

Source: De Bruin, D. (2008). What do you mean your test is cross-culturally valid? *Workshop presented at SIOPSA, Pretoria, SA.*

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Variance estimator of DTF for the scale with all 15 items included

DTF STATISTICS: DICHOTOMOUS ITEMS

Statistic	Value	SE	Z
Tau^2	0.214	0.084	2.548
Weighted Tau^2	0.208	0.081	2.568

With all items included the variance estimator of DTF is 0.214. This may be classified as large DTF (Tau^2 > 0.14).

Source: De Bruin, D. (2008). What do you mean your test is cross-culturally valid? *Workshop presented at SIOPSA, Pretoria, SA.*

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Variance estimator of DTF for the scale with 6 DIF items excluded

DTF STATISTICS: DICHOTOMOUS ITEMS

Statistic	Value	SE	Z
Tau^2	0.022	0.017	1.294
Weighted Tau^2	0.010	0.011	0.909

With six DIF items excluded the variance estimator of DTF is 0.022. This appears to be small to negligible DTF (Tau^2 < 0.07). The reduced scale exhibits very little bias from a statistical perspective, but does the scale still measure what we want it to measure?

Source: De Bruin, D. (2008). What do you mean your test is cross-culturally valid? *Workshop presented at SIOPSA, Pretoria, SA.*

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Extending the MH statistic to polytomous items

- Mantel's (1963) chi-square test (not an extension of the MH test) can be used with polytomous items
- Liu and Agresti (1996) extended the MH statistic for use with ordinal variables
  - The Liu Agresti estimator is a generalization of the MH common odds ratio
- Penfield and Algina (2003) applied the Liu Agresti estimator to detect DIF in polytomous items
  - They provide computational detail
- The Liu Agresti estimator will give similar results as the Mantel test, but has the advantage that it is interpreted in the same frame of reference as the MH common odds ratio

Back to the console...

MH METHOD WITH R

R package for DIF analysis (difR)

- difR is a package that provides several opportunities to calculate *dichotomous* DIF
- it is connected to the ltm package with has also be installed but no calls on that have to be made (all done by difR)
- Reference: Magis, D., Béland, S., Tuerlinckx, F., & Boeck, P. de (2010). *Behavior Research Methods, 42, 847-862*.

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

R package for DIF analysis (difR)

- difR is an R package that provides several procedures to calculate *dichotomous* DIF
- Mantel-Haenszel procedure

```
difMH(Data, group, focal.name , MHstat="MHChisq",
  correct=TRUE, alpha=0.05, purify=FALSE,
  nrIter=10)
```

  - Needs a grouping vector
  - Needs the code of Focal group
  - Needs an object containing the items

Example: PROMIS Data

The data contain responses given by 766 people sampled from a general population to the PROMIS Anxiety scale (`http://www.nihpromis.org`) composed of 29 Likert-type questions with a common rating scale (1=Never, 2=Rarely, 3=Sometimes, 4=Often, and 5=Always).

`age` 0=younger than 65 and 1=65 and older

`gender` 0=Male and 1=Female

`education` 0=some college or higher and 1=high school or lower

`R1` I felt fearful

`R2` I felt frightened

`R3` It scared me when I felt nervous

`R4` I felt anxious

`R5` I felt like I needed help for my anxiety

`R6` I was concerned about my mental health

- From Choi (2011): *lordif* manual on CRAN.

Mantel-Haenszel in difR

- Import the data file (dichotomized (0-0-1-1-1) version of PROMIS Anxiety Data)

```
PROMIS2cat <-
  read.table(file.choose(),
  header=TRUE, sep="\t",
  na.strings="NA", dec=".",
  strip.white=TRUE)
```

- activate "difR"

```
library(difR)
```

Mantel-Haenszel in difR

- options of the "difMH" function (handbook)

```
difMH(Data, group, focal.name , MHstat="MHChisq",
    correct=TRUE, alpha=0.05, purify=FALSE, nrIter=10,
    save.output=FALSE, output=c("out","default"))
```

- Create grouping variable and item set

```
age2cat<-PROMIS2cat[ ,1]
items<-PROMIS2cat[ ,4:32]
```

- Call the MH function

```
resMH1<-difMH(items,age2cat,focal.name=0)
```

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Mantel-Haenszel in difR

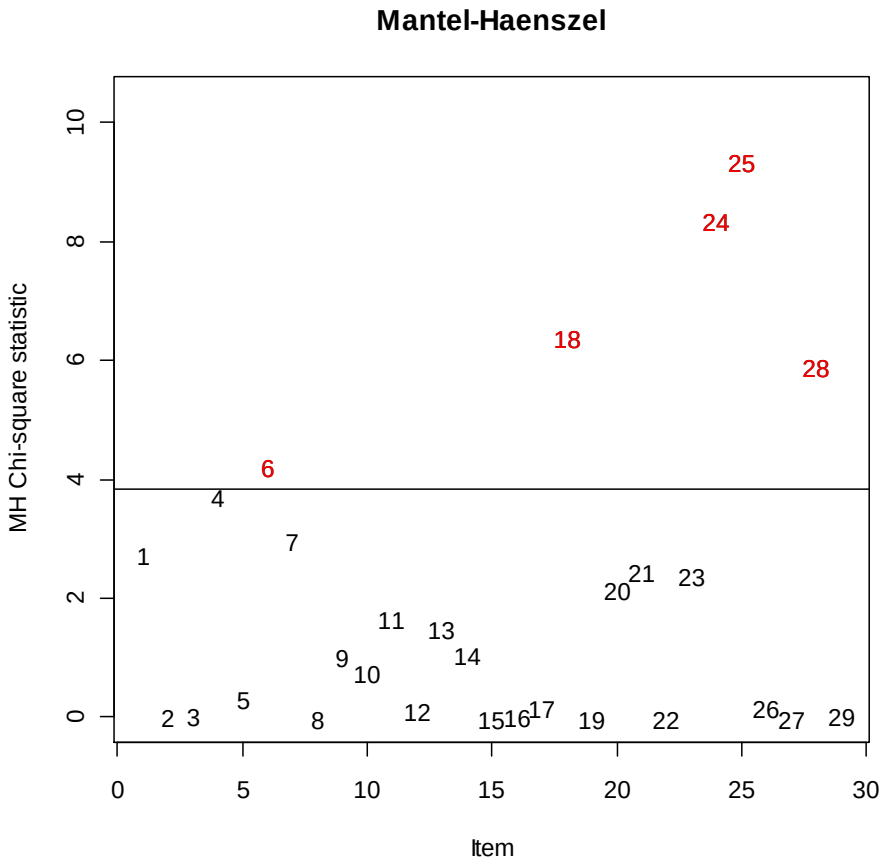
```
resMH1

Mantel-Haenszel Chi-square statistic:

     Stat.  P-value
R1   2.7416 0.0978  .
R2   0.0243 0.8761
R3   0.0409 0.8398
R4   3.7278 0.0535  .
R5   0.3191 0.5721
R6   4.2181 0.0400  *
R7   2.9736 0.0846  .
```

- overall five items show significant DIF with respect to AGE

Mantel-Haenszel in difR

- You can also ask for purification
- Results can be plotted

  plot(resMH1)



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Mantel-Haenszel chi-square (for previous example)

- Results from R

Data Editor

	resMH1.resMH
1	0.2460909
2	7.658029
3	1.816225
4	32.46579
5	0.03419603
6	82.82315
7	0.3814386
8	0.6644087
9	4.906731
10	31.23271
11	5.859884
12	33.04937
13	1.957509
14	5.079773
15	24.69691
16	

and from DIFAS

```
DIF STATISTICS: DICHO
----------------------
Name              MH CHI
----------------------
Var 1              0.2461
Var 2               7.658
Var 3              1.8162
Var 4             32.4658
Var 5              0.0342
Var 6             82.8232
Var 7              0.3814
Var 8              0.6644
Var 9              4.9067
Var 10            31.2327
Var 11             5.8599
Var 12            33.0494
Var 13             1.9575
Var 14             5.0798
Var 15            24.6969
----------------------
```


Practical

- Please test for DIF in the other two grouping variables with difR

Interpreting DIF

- Should we be driven by statistical or practical significance?
- Certainly the most important consideration is the impact of DIF on the test score
  - This is why DTF is important
  - When the test is not fixed (e.g. randomised), DTF cannot be computed
  - Then compute the impact of this item on the test score
- Remember that DIF studies are only precursor to item bias studies
  - Advice from Ron Hambleton: arrange the items in the order of DIF magnitude and start interpreting
  - When cannot interpret DIF anymore, stop

How to deal with DIF

- If an item is demonstrating DIF, do not immediately get rid of it
  - The domain being tapped will become too limited quickly
  - Reliability might be compromised
  - Further studies might be required
  - Final decision will depend on the impact
- In test adaptation
  - Non-equivalent items across the intended populations should not be used in “linking” adapted version of the test to a common scale.
  - However, these same items may be useful for reporting scores in each population separately.

How to adjust for DIF

- It is also possible to adjust for DIF in the model
  - For example, can add direct effect between the group and the item in M*plus*
- Crane et al. (2004, 2006)

a) items without DIF have item parameters estimated from whole sample – (anchors)

b) items with DIF have parameters estimated separately in different subgroups

Item purification (e.g. Magis et al., 2010)

1. Test all items one by one, assuming they are not DIF items.
2. Define a set of DIF items on the basis of the results of Step 1.
3. If the set of DIF items is empty after the first iteration, or if this set is identical to the one obtained in the previous iteration, then go to Step 6. Otherwise, go to Step 4.
4. Test all items one by one, omitting the items from the set obtained in Step 2, except when the DIF item in question is being tested.
5. Define a set of DIF items on the basis of the results of Step 4 and go to Step 3.
6. Stop.

Mantel-Haenszel in difR

- "scale purification" is an automated option
- only items without DIF will be used for stratification:

```
resMH2<-difMH(items,age2cat,
focal.name=0, purify=TRUE)
```

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Mantel-Haenszel in difR

- classification of DIF size:

```
Effect size code:
 'A': negligible effect
 'B': moderate effect
 'C': large effect

R1    2.6119 -2.2561 C
R2    1.3791 -0.7554 A
R3    1.5593 -1.0439 B
R4    2.5005 -2.1537 C
R6    0.4543  1.8541 C
R10   0.5204  1.5347 C
R18   0.5063  1.5994 C
R24   2.6194 -2.2629 C
R25   2.0397 -1.6751 C
R28   0.4835  1.7078 C
```

several items show in size considerable DIF with regard to the AGE group

MORE DIF IN R

Two methods for DIF

- methods relying on an IRT-model (i.e. IRT-methods, [„parametric methods“])

- methods NOT relying on an IRT model (i.e. non-IRT methods, [„nonparametric“])

Two methods for DIF (Magis et al., 2010)

Traditional Methods for Detecting Differential Item Functioning (DIF)

		Number of Groups	
Framework	DIF Effect	2	>2
Non-IRT	Uniform	Mantel–Haenszel*	Pairwise comparisons
		Standardization*	Generalized Mantel–Haenszel*
		SIBTEST	
		Logistic regression*	
Non-IRT	Nonuniform	Logistic regression*	Pairwise comparisons
		Breslow–Day*	
		NU.MH	
		NU.SIBTEST	
IRT	Uniform	LRT*	Pairwise comparisons
		Lord*	Generalized Lord*
		Raju*	
IRT	Nonuniform	LRT*	Pairwise comparisons
		Lord*	Generalized Lord*
		Raju*	

Note—NU.MH, modified Mantel–Haenszel for nonuniform DIF; NU.SIBTEST, modified SIBTEST for nonuniform DIF; LRT, likelihood ratio test. *Currently implemented in difR package (Version 2.2).

DIF in difR: Breslow Day

- MH for uniform DIF was already discussed
- for non-uniform DIF: Breslow-Day
- determines whether the association between item response and group membership is homogeneous over the range of the scale

$$\mathrm{BD} = \sum_j \frac{\left[A_j - \mathrm{E}\left(A_j\right)\right]^2}{\mathrm{Var}\left(A_j\right)}.$$

- defined with A being the number of correct responses and ist expectation based on the odds-ratio between groups

DIF in difR: Breslow Day

- Breslow Day statistic and its options in difR:

```
difBD(Data, group, focal.name, BDstat="BD", alpha=0.05,
   purify=FALSE, nrIter=10, save.output=FALSE,
   output=c("out","default"))
```

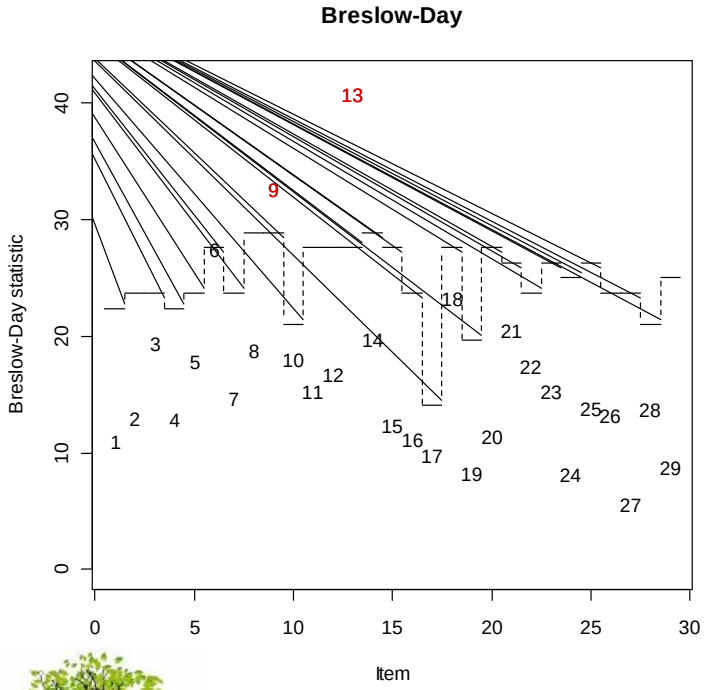
```
resBD<-difBD(items,age2cat,
focal.name=0, purify=TRUE,
  nrIter=150)
```

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

DIF in difR: Breslow Day

Results in our example indicate only one item that probably shows non-uniform DIF:

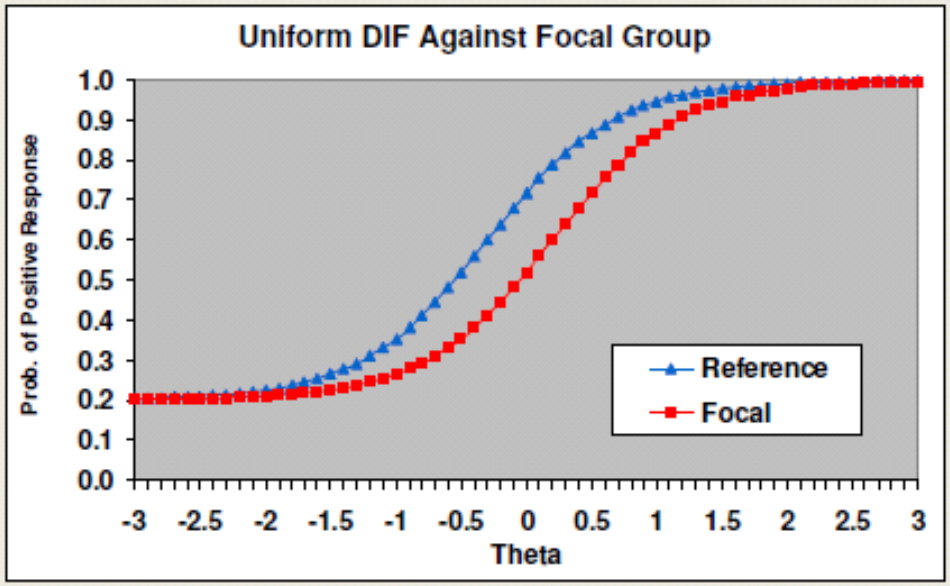
```
Breslow-Day statistic:

     Stat.    df       P-value
R6   27.5485 17.0000   0.0505 .
R9   32.7180 18.0000   0.0181 *
R13  40.9178 17.0000   0.0010 ***
```



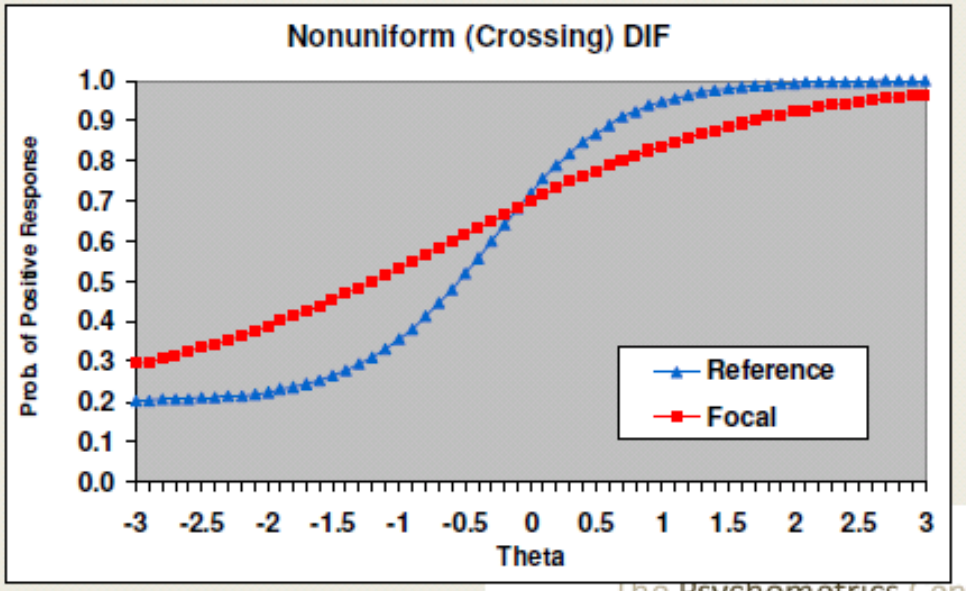
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

LOGISTIC REGRESSION TO DETECT DIF

Uniform and non-uniform DIF

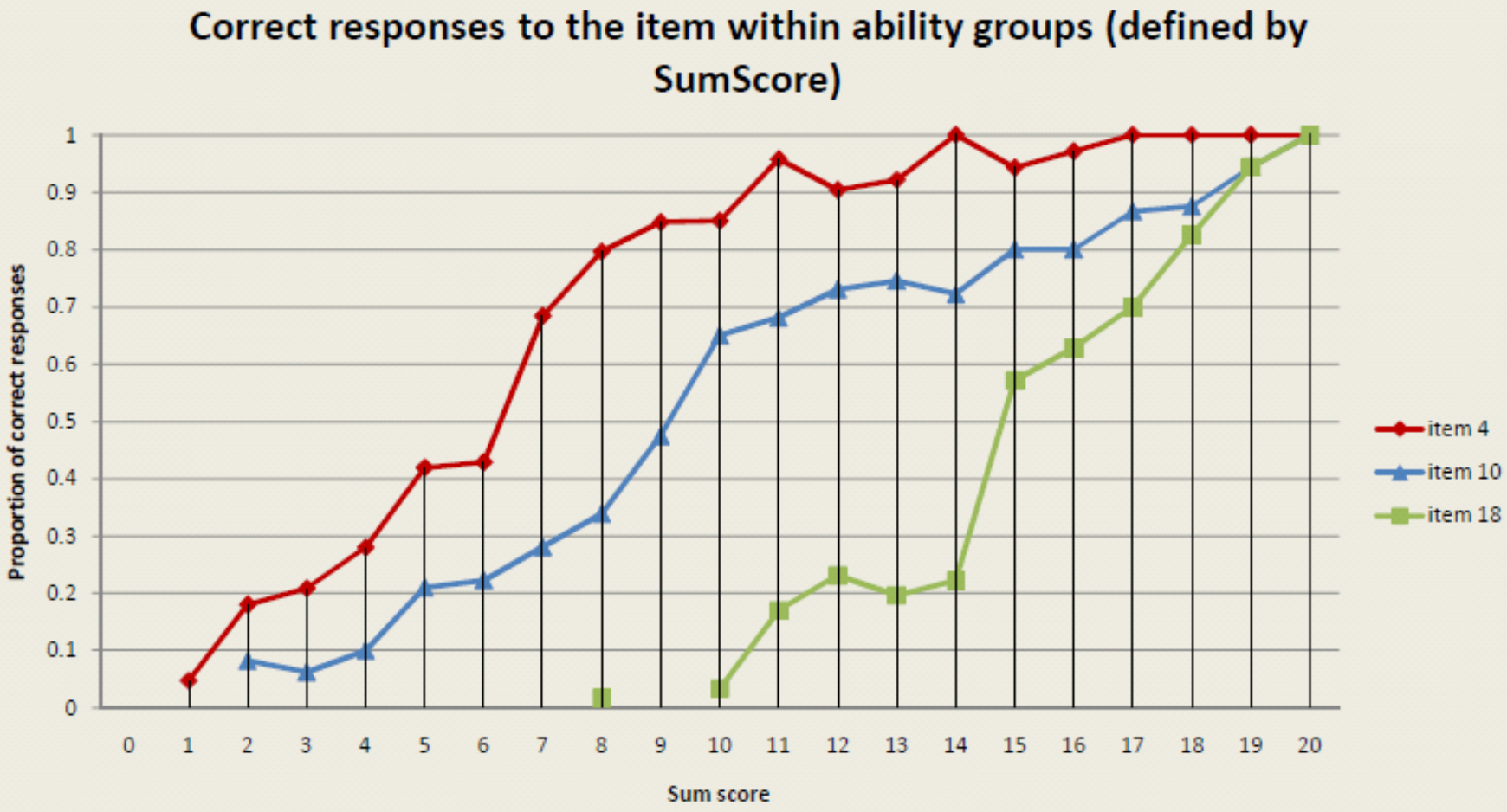


Focal group has lower probability of endorsing the item at all trait levels



Focal group has higher probability of endorsing the item at low level of trait, but lower probability at high level

What can be said about these items?



Logistic Regression to detect DIF

- it is assumed that you have a representation of the latent construct
  - sum score, estimate of ability from IRT model…
- empirical relative frequencies of endorsing an item depending on this proxy for the latent construct should show an approximation of the item characteristic curve

Logistic Regression to detect DIF

- we run a logistic regression that predicts the probability to solve an item from the level of the latent construct

$$P(X_{vi}=1)=\frac{e^{a+b_1 trait+b_2 grouping}}{1+e^{a+b_1 trait+b_2 grouping}}$$

$$\ln\left(\frac{P(X_{vi}=1)}{1-P(X_{vi}=1)}\right)=a+b_1 trait+b_2 grouping$$

Logistic Regression to detect DIF

- this should (re-)produce our well known ICC
- if we use group in this regression, it is only significant in case of UNIFORM DIF

$$P(X_{vi}=1)=\frac{e^{a+b_1 trait+b_2 grouping}}{1+e^{a+b_1 trait+b_2 grouping}}$$

$$\ln\left(\frac{P(X_{vi}=1)}{1-P(X_{vi}=1)}\right)=a+b_1 trait+b_2 grouping$$

Logistic Regression to detect DIF

- if no uniform DIF was present, the test whether the grouping variable explains additional information beyond the score should not be significant

$$P(X_{vi}=1)=\frac{e^{a+b_1 trait+b_2 grouping}}{1+e^{a+b_1 trait+b_2 grouping}}$$

$$\ln\left(\frac{P(X_{vi}=1)}{1-P(X_{vi}=1)}\right)=a+b_1 trait+b_2 grouping$$

Logistic Regression to detect DIF

- Logistic regression with *score* as predictor

```
RegrUDIF<-
   difLogistic(items,age2cat,criterion="LRT",
   type="udif",alpha=.01,purify=TRUE,
   focal.name=1, nrIter=50)
```

command criterion: LRT, Wald

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Logistic Regression to detect DIF

- As a first result a table with all model tests is provided:

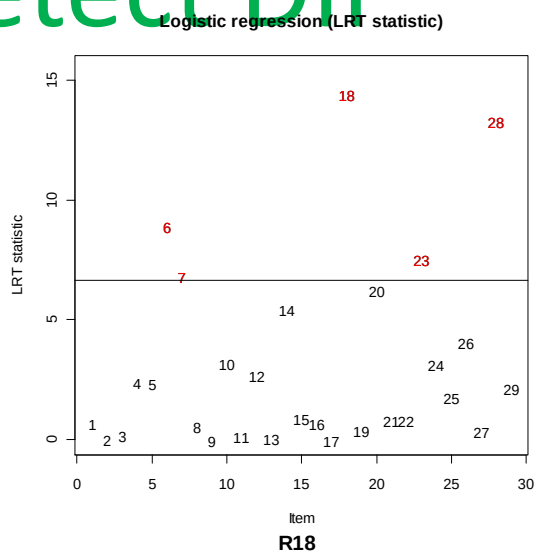
```
Logistic regression DIF statistic:

    Stat.   P-value
R6   8.8881  0.0029 **
R7   6.8168  0.0090 **
R14  5.4704  0.0193 *
R18 14.3925  0.0001 ***
R20  6.2390  0.0125 *
R23  7.5538  0.0060 **
R26  4.0620  0.0439 *
R28 13.2631  0.0003 ***

Detection threshold: 6.6349 (significance level: 0.01)
```

- here all p < .05 coefficients are displayed
- detection threshold is lower
- two or three candidate items identified

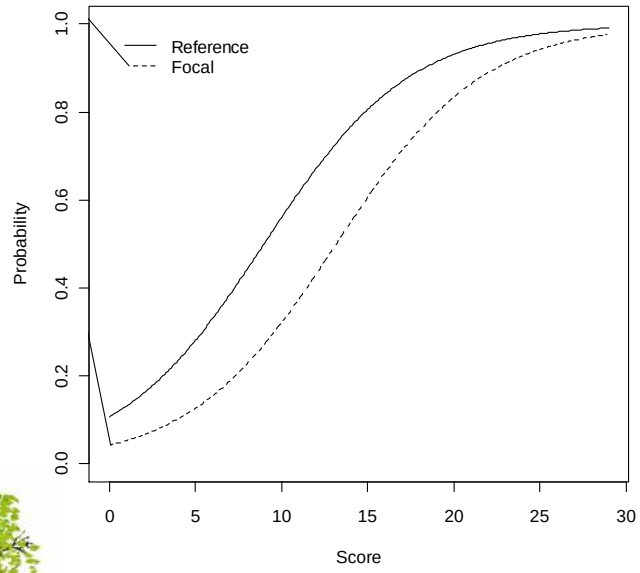
Logistic Regression to detect DIF

- plotting:

```
plot(RegrUDIF)
```




```
plot(dichUDIF,
  plot=
  "itemCurve",
  item=18)
```



Logistic Regression to detect DIF

- usually identification based on effect size:
  - <.035 negligible
  - <.07 moderate
  - >.07 large
- at least according to this criterion: no uniform DIF effect for age

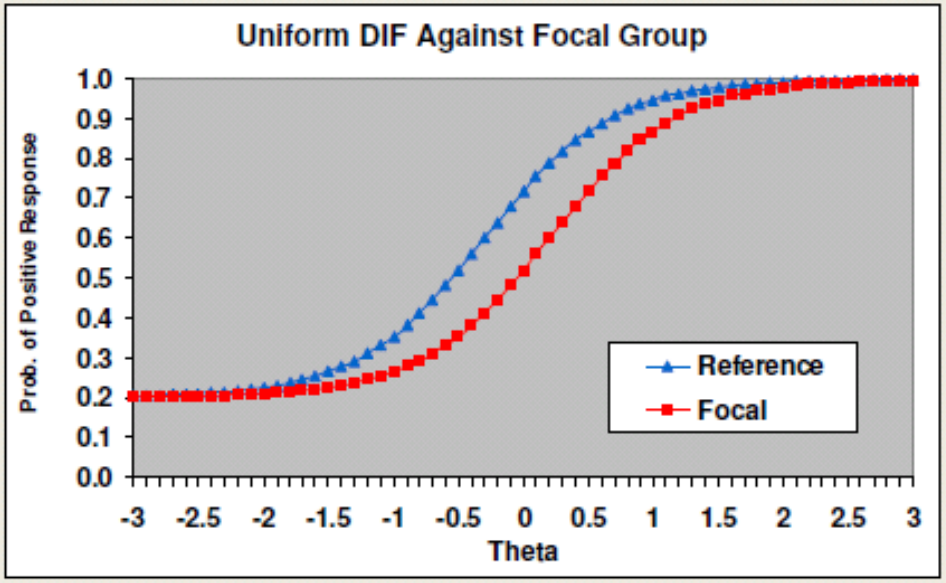
(Crane et al., 2007)

```
Effect size code:
 'A': negligible
  effect
 'B': moderate
  effect
 'C': large effect
```

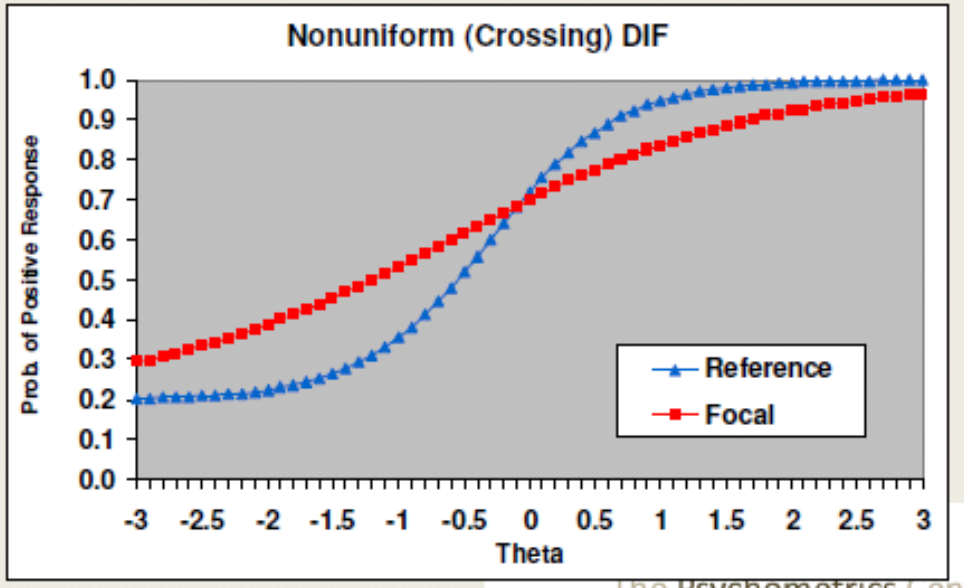
	R^2	ZT	JG
R1	0.0015	A	A
R2	0.0001	A	A
R3	0.0004	A	A
R4	0.0033	A	A
R5	0.0050	A	A

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Uniform and non-uniform DIF



Focal group has lower probability of endorsing the item at all trait levels



Focal group has higher probability of endorsing the item at low level of trait, but lower probability at high level

Logistic Regression to detect DIF

- non-uniform DIF adds to this only the interaction between grouping (G) and the construct level (T)

$$P(X_{vi}=1)=\frac{e^{a+b_1 trait+b_2 grouping+b_3 TG}}{1+e^{a+b_1 trait+b_2 grouping+b_3 TG}}$$

$$\ln\left(\frac{P(X_{vi}=1)}{1-P(X_{vi}=1)}\right)=a+b_1 trait+b_2 grouping+b_3 TG$$

Logistic Regression to detect DIF

- if this interaction terms adds significant as well as relevant information compared to the uniform DIF, it is flagged

$$P(X_{vi}=1)=\frac{e^{a+b_1 trait+b_2 grouping+b_3 TG}}{1+e^{a+b_1 trait+b_2 grouping+b_3 TG}}$$

$$\ln\left(\frac{P(X_{vi}=1)}{1-P(X_{vi}=1)}\right)=a+b_1 trait+b_2 grouping+b_3 TG$$

Logistic Regression to detect DIF

```
NonUDIFRegr<-
  difLogistic(items,age2cat,criterion="LRT",
  type="nudif",alpha=.01,purify=TRUE,
  focal.name=1)
```

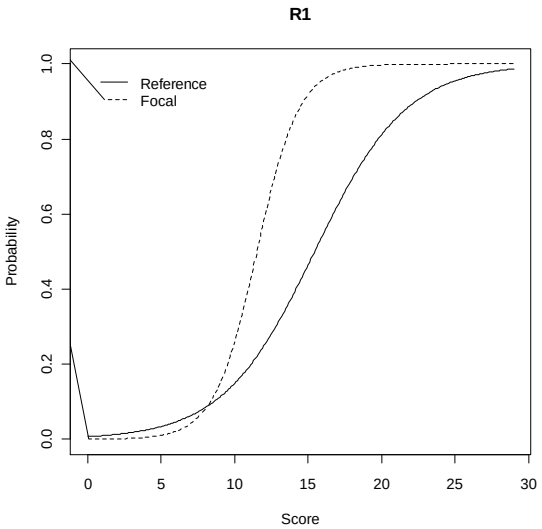
Logistic Regression to detect DIF

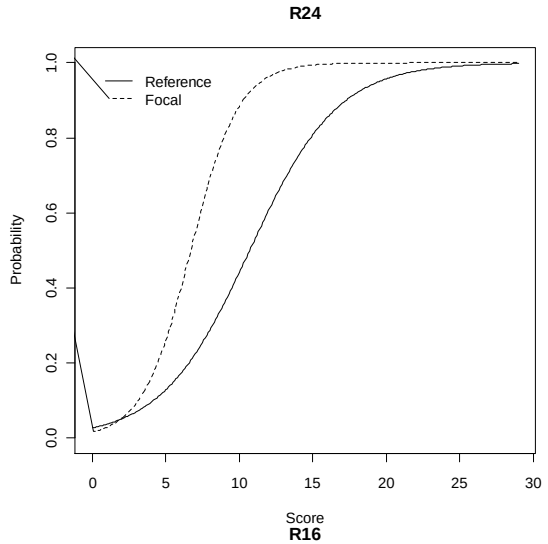
- Again display of the items with possible DIF effects according to significance; if predefined detection level is used, five items can be suspecte dto show non-uniform DIF:

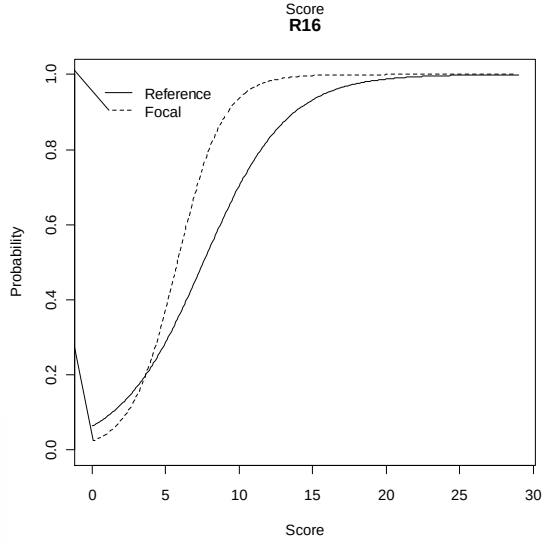
```
    Stat.  P-value
R1  8.9150 0.0028  **
R2  6.9591 0.0083  **
R9  6.6204 0.0101  *
R16 9.6509 0.0019  **
R24 9.7432 0.0018  **
R25 7.8684 0.0050  **
R26 3.8618 0.0494  *

Detection threshold: 6.6349 (significance level: 0.01)
```

Logistic Regression to detect DIF

- again none of the effect size criteria show a relevant DIF effect



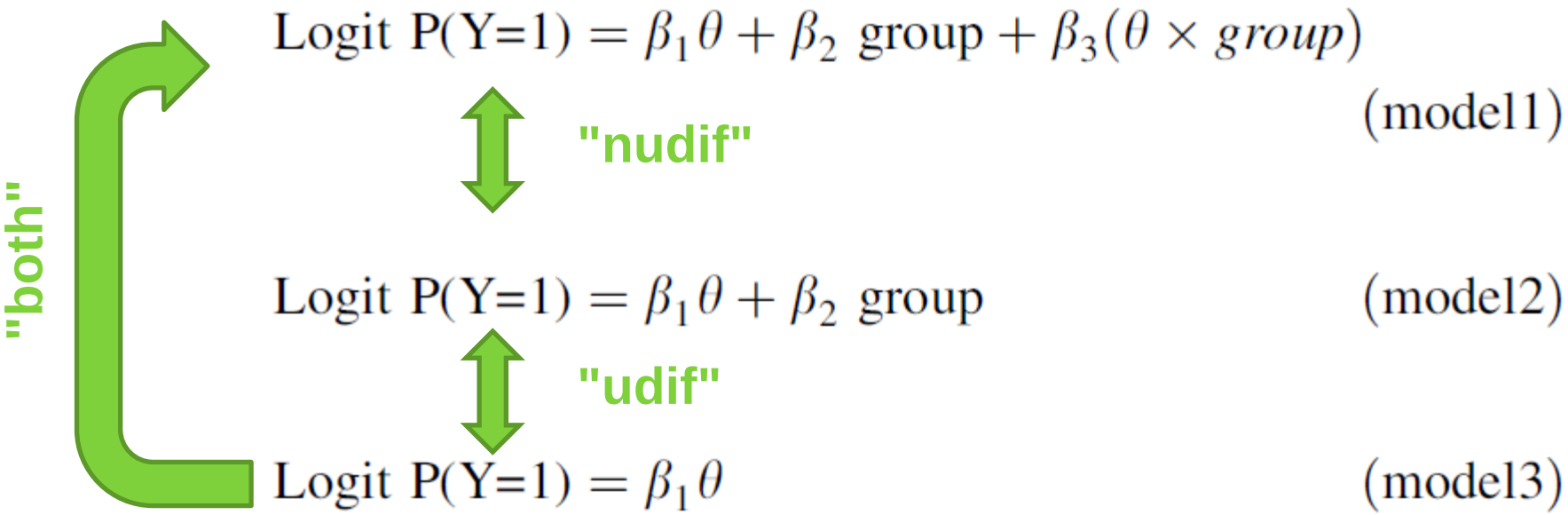





```
plot(NonUDIFRegr,plot=
   "itemCurve",item=1)
```

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

DIF in difR



Crane et al. (2007). Qual Life Res, 16, 69-84.

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Logistic Regression to detect DIF

```
allDIFRegr<-
  difLogistic(items,age2cat,criterion="LRT",
  type="both",alpha=.01,purify=TRUE,
  focal.name=1)
```

This test for whether items might sow DIF at all, whether it be of uniform or non-uniform

Logistic Regression to detect DIF

- Again display of the items with possible DIF effects according to significance; if predefined detection level is used, six items can be suspecte dto show any DIF:

```
    Stat.  P-value
 Stat.   P-value
R1    9.7874  0.0075 **
R4    6.2519  0.0439 *
R6    7.6747  0.0216 *
R9    7.3476  0.0254 *
R16 12.1178  0.0023 **
R18 13.2328  0.0013 **
R24 16.8380  0.0002 ***
R25 11.0070  0.0041 **
R26  7.9750  0.0185 *
R28 12.8000  0.0017 **

Detection threshold: 6.6349 (significance level: 0.01)
```

Probing multiple criteria

- difR provides the opportunity to use multiple criteria at the same time instead of tediously one after another:

```
generalDIF<-
  dichoDif(items,age2cat,focal.na
  me=1, method=c("MH","Logistic",
  "BD"), alpha=.01,purify=TRUE,
  nrIter=100)
```

Comparison

- Logistic regression tests the IRT hypothesis: that there is something like the Item Characteristic Curve linking latent construct and probability for the specific response
- but only when the representation of the latent construct is correct!
- Mantel-Haenszel less prone to this error

Practical

- Please use difR's logistic reression to test for DIF with respect to gender and education!

POLYTOMOUS ITEMS

MH polytomous

- only implemented by the mantelhaen.test() command
  - here the score has to be defined
  - and it has to be done for every item by hand…

- difR only for dichotomous items (yet at least...)

Logistic Regression polytomous

```
library(lordif)
```

- this package contains the command lordif()
- estimates the latent construct via Graded Response Model (from ltm)
- conditions in this case on estimated thetas and not on the score
- purification is always performed

Logistic Regression polytomous

- implemented in the „lordif“ package
- polytomous data in package:

The data contain responses given by 766 people sampled from a general population to the PROMIS Anxiety scale (`http://www.nihpromis.org`) composed of 29 Likert-type questions with a common rating scale (1=Never, 2=Rarely, 3=Sometimes, 4=Often, and 5=Always).

- **read data:** `data(Anxiety)`

Logistic Regression polytomous

The data contain responses given by 766 people sampled from a general population to the PROMIS Anxiety scale (http://www.nihpromis.org) composed of 29 Likert-type questions with a common rating scale (1=Never, 2=Rarely, 3=Sometimes, 4=Often, and 5=Always).

- first three items contain demographics
- define only items:

```
Anxiety.poly<-Anxiety[ ,4:19]
head(Anxiety.poly)
```

- define grouping vector, e.g. "age" (=1st variable):

```
age<-Anxiety[ ,1]
```

Logistic Regression polytomous

- Building code for our example:

```
lordif(resp.data, group, selection = NULL,
criterion = c("Chisqr", "R2", "Beta"),
pseudo.R2 = c("McFadden", "Nagelkerke", "CoxSnell"), alpha = 0.01,
beta.change = 0.1, R2.change = 0.02, maxIter = 10, minCell = 5,
minTheta = -4, maxTheta = 4, inc = 0.1, NQ=41)
```

- regression.test<-lordif(Anxiety.poly, age, criterion="R2")

Assessing results

- R directly tells whether items with DIF were encountered:

```
regression.test<-lordif(Anxiety.poly, age)
Iteration 1 : 4 items flagged for DIF (1,7,9,11)
Iteration 2 : 4 items flagged for DIF (1,7,9,11)
```

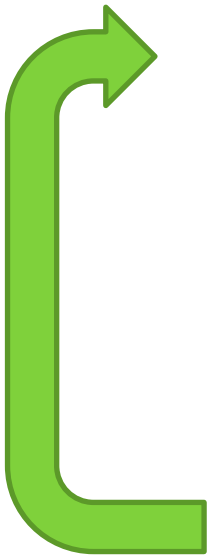
- this run needed 2 iterations, in both the same items were identified as showing DIF

Assessing results

- R report `regression.test`
- R report
  - `chi12` test for uniform DIF compared to baseline model
  - `chi13` test for general DIF compared to baseline model
  - `chi23` test between uniform & non-uniform

```
Number of iterations for purification: 2 of 10

Detection criterion: Chisqr

Threshold: alpha = 0.01

   item ncat  chi12  chi13  chi23
1     1    3 0.8026 0.0022 0.0005
2     2    3 0.1204 0.0240 0.0246
3     3    3 0.6626 0.5346 0.3027
4     4    4 0.2119 0.4332 0.7347
5     5    3 0.3249 0.4381 0.4090
6     6    3 0.2538 0.0327 0.0186
7     7    3 0.0005 0.0018 0.4237
8     8    3 0.8191 0.9684 0.9128
9     9    3 0.0025 0.0069 0.3675
10   10    3 0.0630 0.1775 0.9921
11   11    3 0.0019 0.0036 0.2133
12   12    3 0.1358 0.1108 0.1402
13   13    3 0.8702 0.7101 0.4173
14   14    3 0.3487 0.6383 0.8884
15   15    3 0.2618 0.3973 0.4437
16   16    3 0.5618 0.7280 0.5849
```

DIF in difR



$$\text{Logit } P(Y=1) = \beta_1\theta + \beta_2 \text{ group} + \beta_3(\theta \times group)$$ Model 3

$$\text{Logit } P(Y=1) = \beta_1\theta + \beta_2 \text{ group}$$ Model 2

$$\text{Logit } P(Y=1) = \beta_1\theta$$ Model 1

Crane et al. (2007). Qual Life Res, 16, 69-84.

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

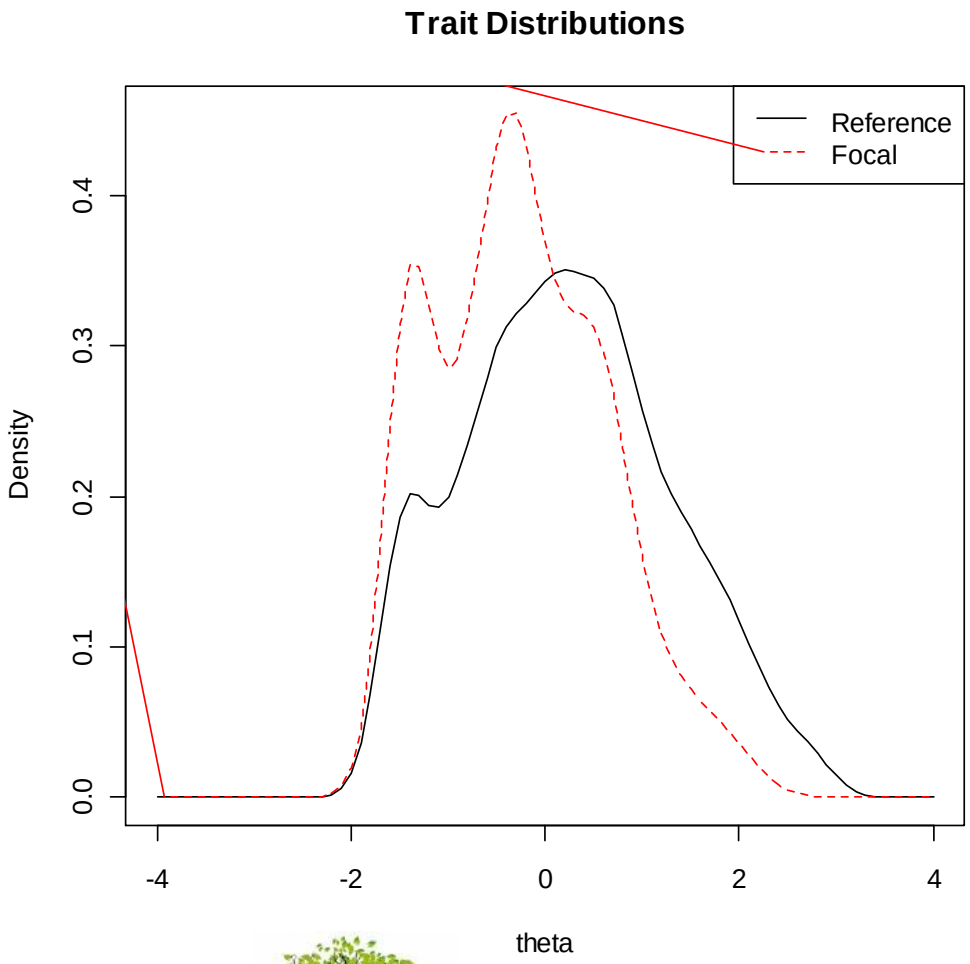
Plot

- plot() will produce a series of plots to evaluate the impact of the DIF items on the current scale

```
plot(regression.test)
```

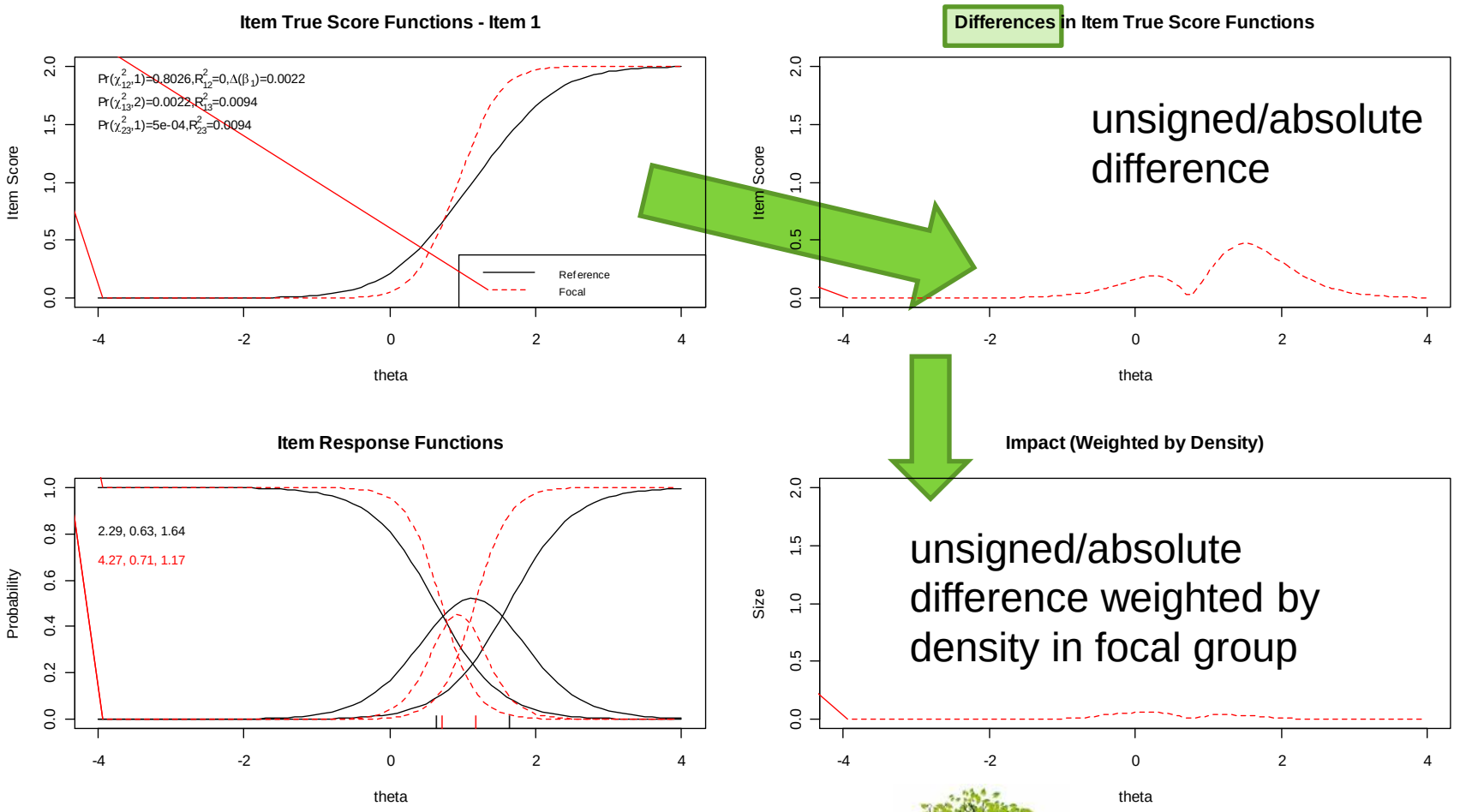
Plot

- latent construct distributions of reference and focal groups
- refrence group is smaller number in group vector, here "younger than 65"



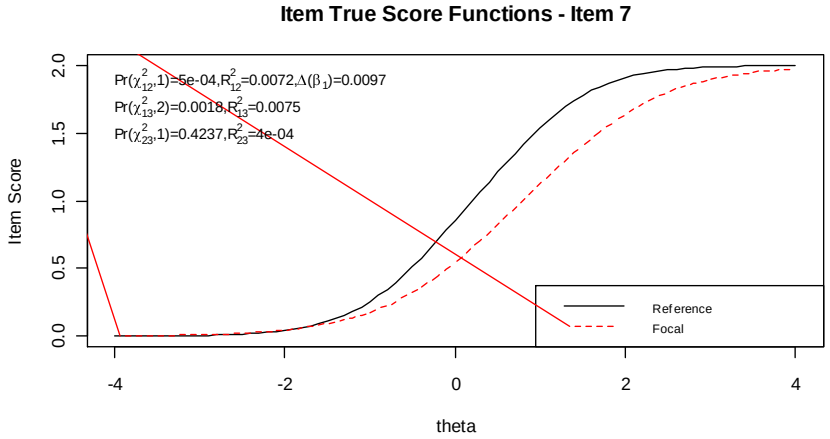
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

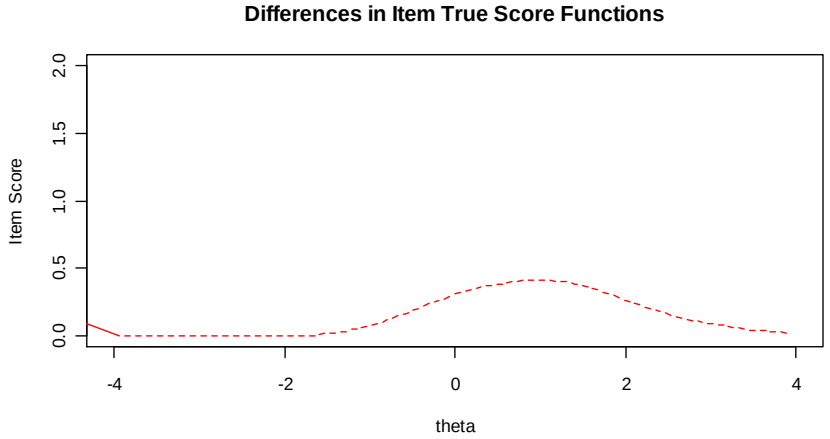
Plot

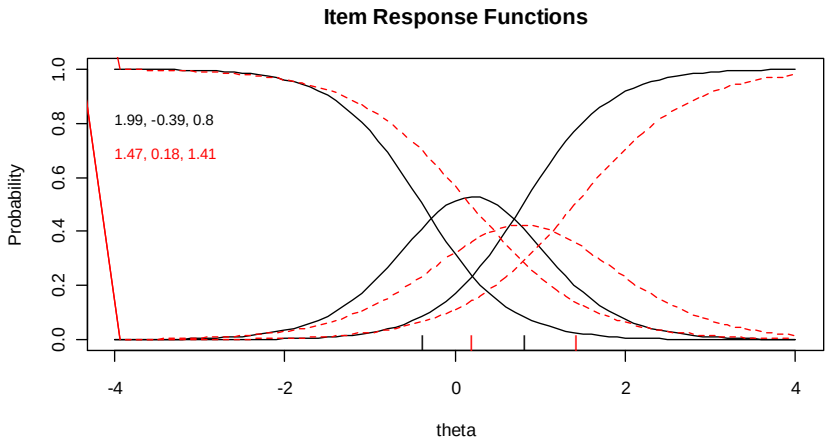


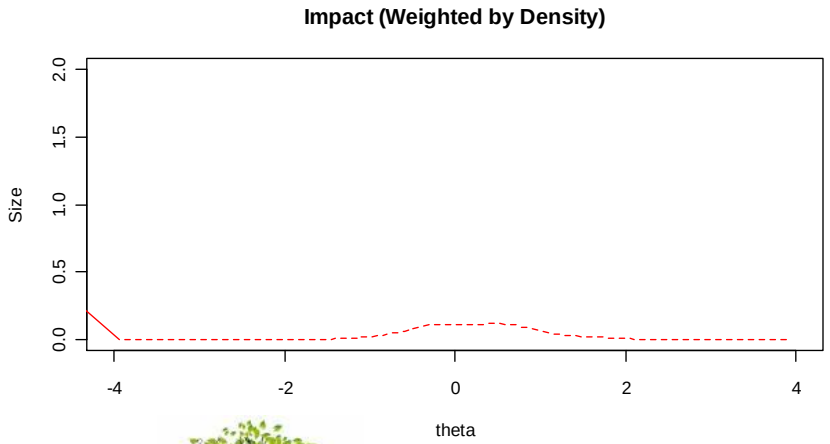
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Plot



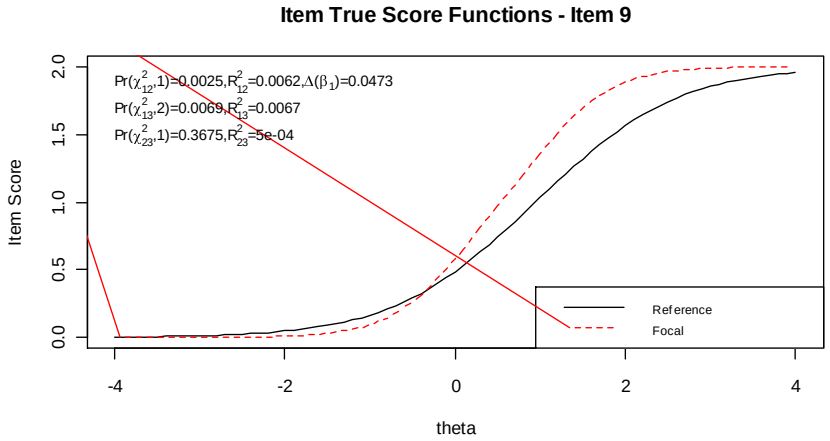


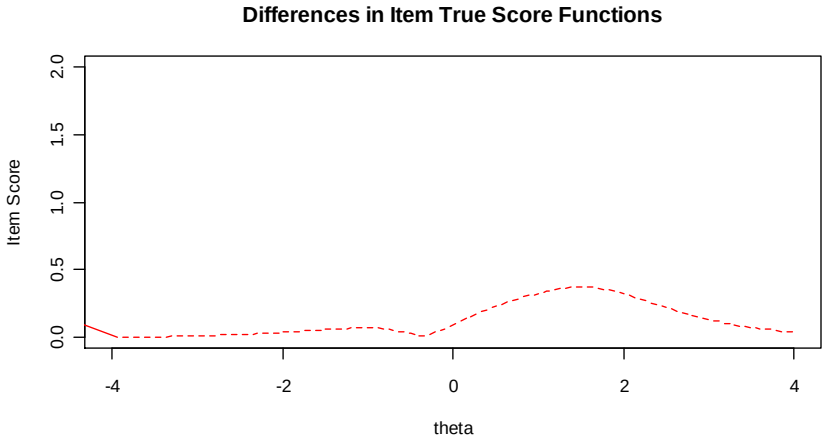


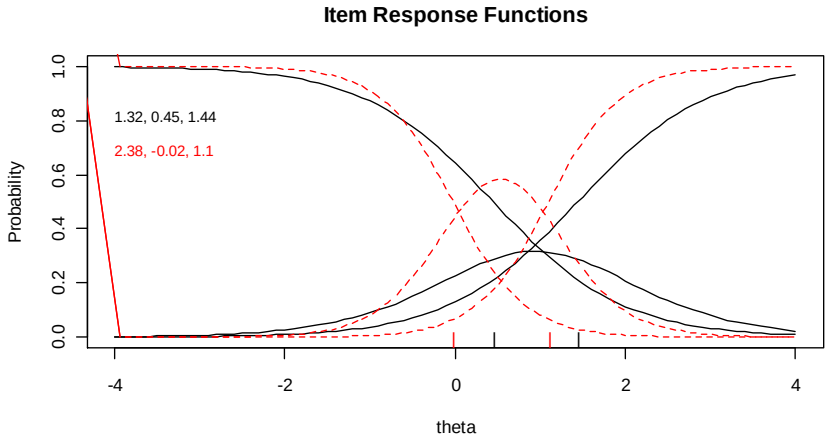


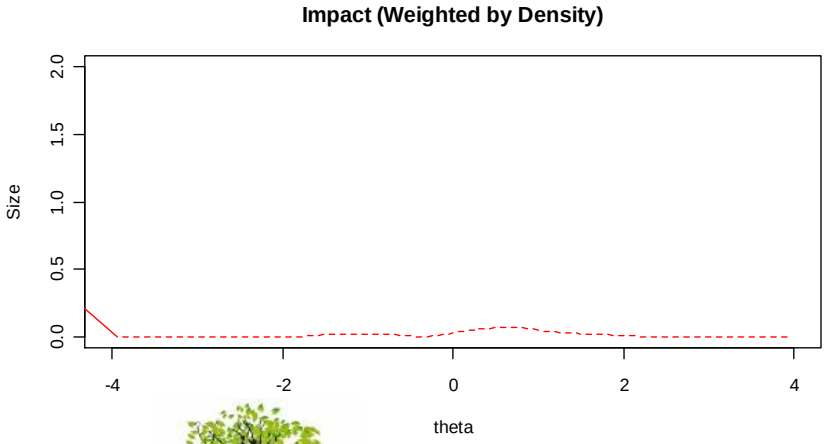
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Plot



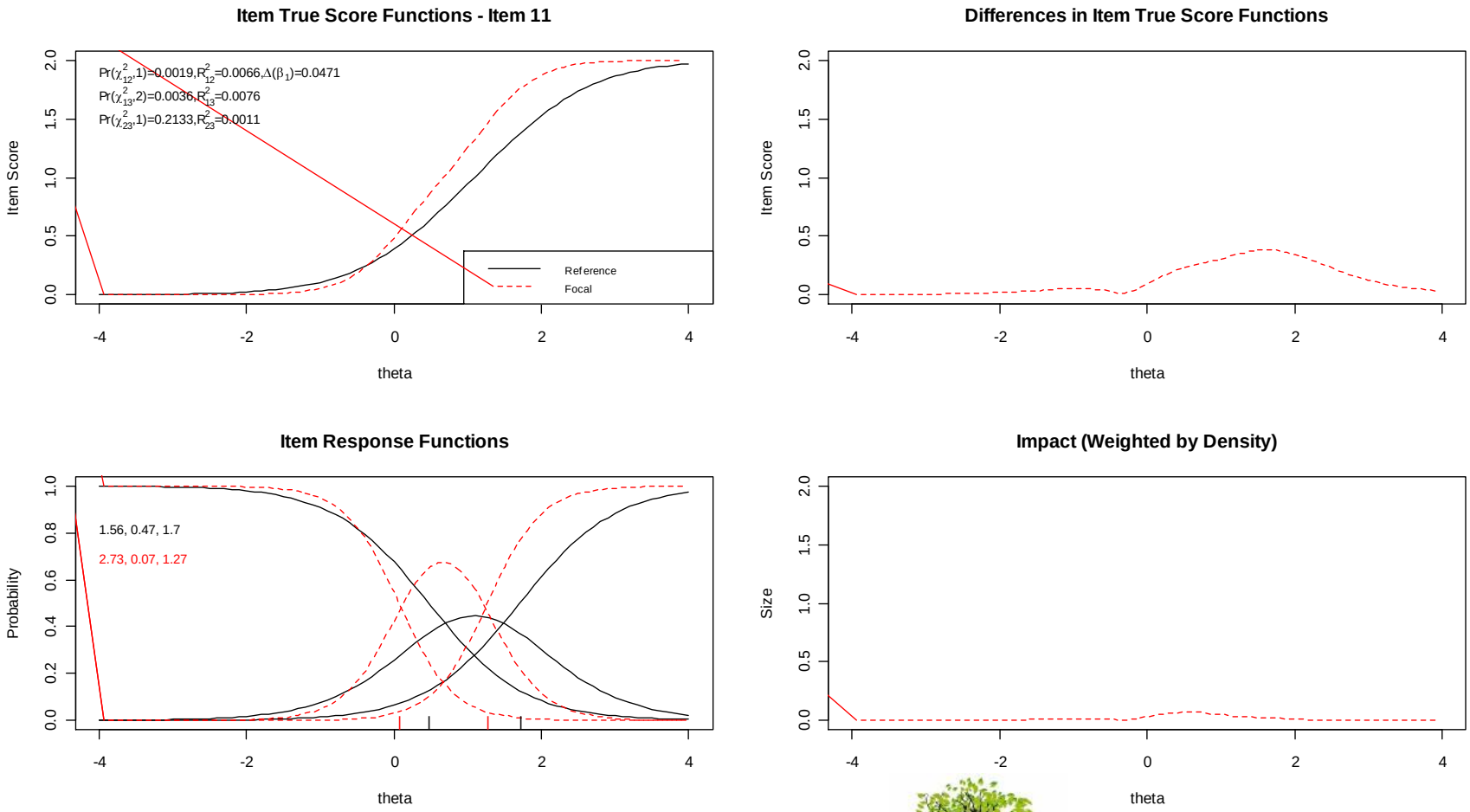




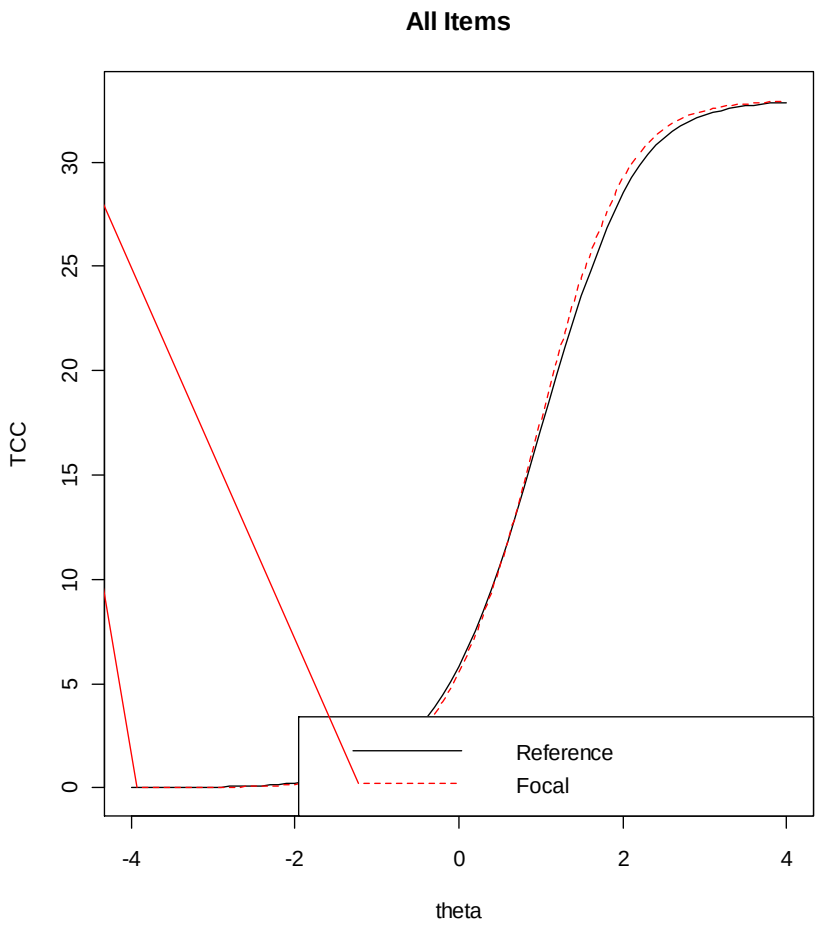


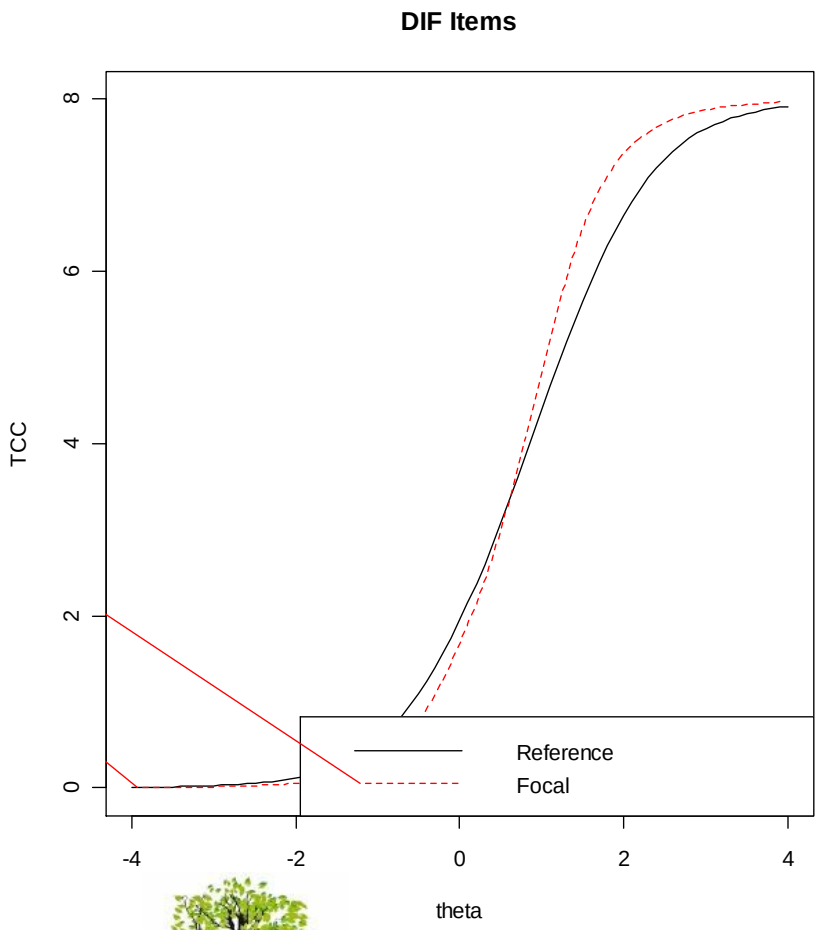
UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Plot


Item True Score Functions - Item 11
Pr(χ²12,1)=0.0019,R²12=0.0066,Λ(β1)=0.0471
Pr(χ²13,2)=0.0036,R²13=0.0076
Pr(χ²23,1)=0.2133,R²23=0.0011
Reference
Focal
Item Score
theta
Differences in Item True Score Functions
Item Score
theta
Item Response Functions
1.56, 0.47, 1.7
2.73, 0.07, 1.27
Probability
theta
Impact (Weighted by Density)
Size
theta

Plot of test characteristc curves





UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Practical

- Please test with lordif() for DIF with respect to gender and age!

Monte Carlo Testing

- Since this is also highly dependent on distributional assumptions, a Monte Carlo Procedure is implemented to be less bound to these assumptions

Monte Carlo Cut Offs

- The routine montecarlo() generates B sets of simulated data
- these are generated under the conditions of
  - IRT model (GRM) is true
  - and no DIF is present
- from these cut off values can be inferred that can be used to identify under the specific conditions in the data set

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

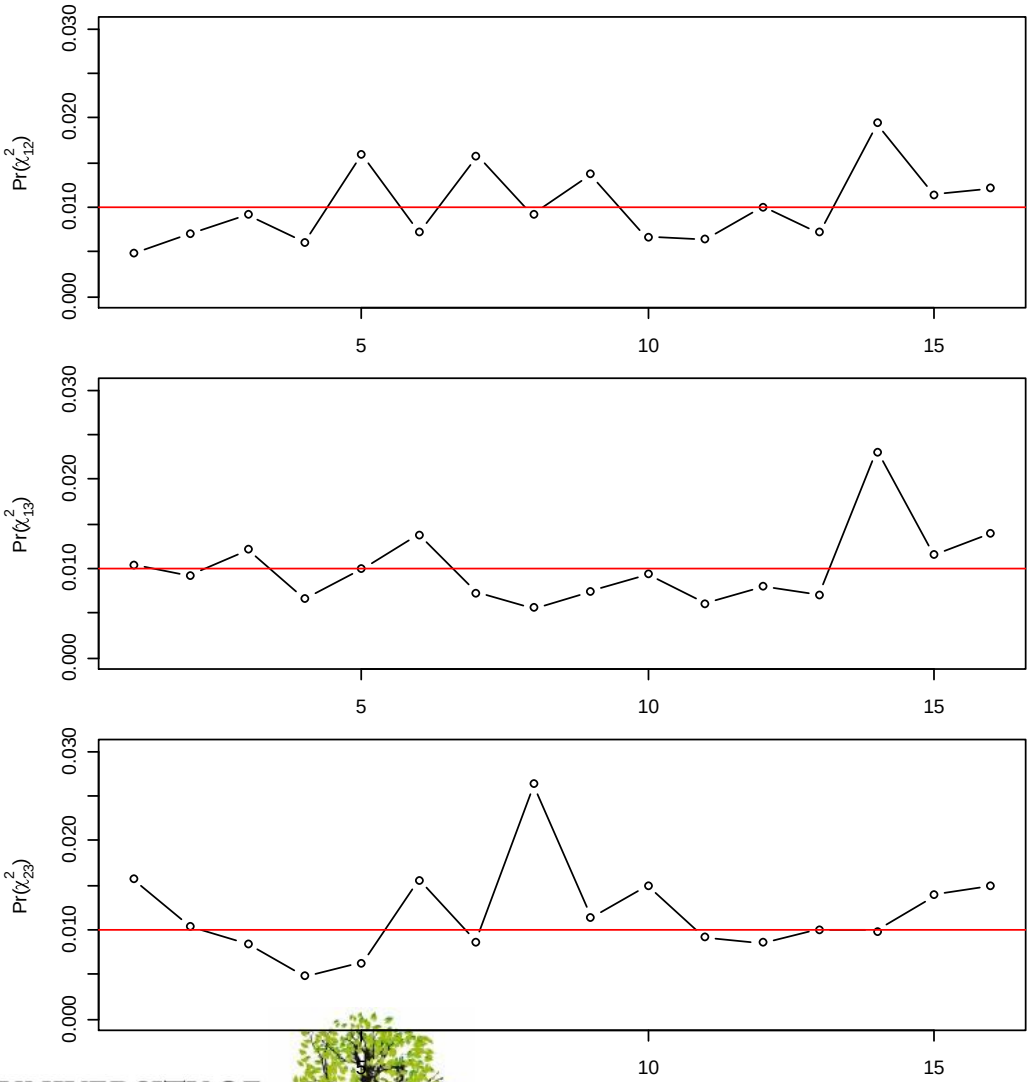
Monte Carlo Cut Offs

- Syntax for this fairly easy: used on the just estimated model; only the number of simulated samples has to be specified

```
montecarlo.DIF<-montecarlo(regression.test,
   nr=500)

plot.lordif.MC(montecarlo.DIF)
```

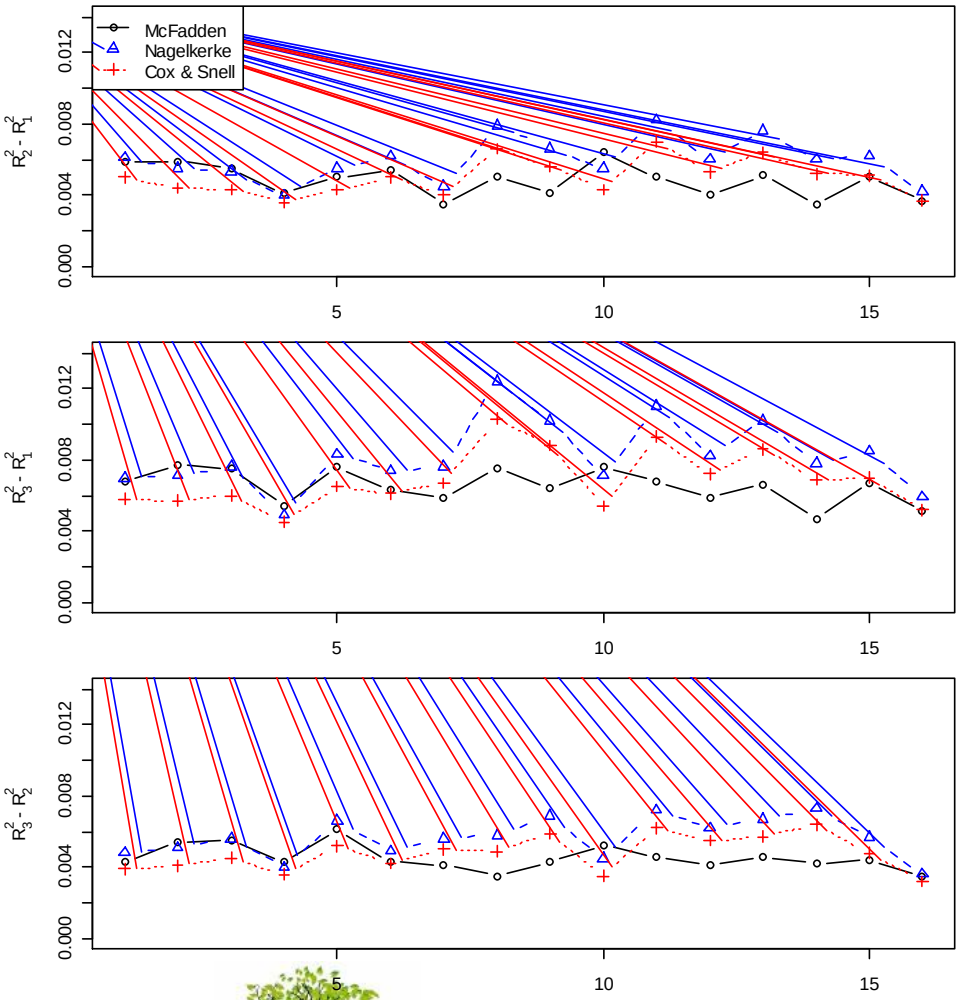
Monte Carlo Cut Offs

- x-Axis: Itemnr
- red = nominal alpha level
- black = actual probability for that specific item
- usually p < .05 is used; debate over adjustment for repeated testing



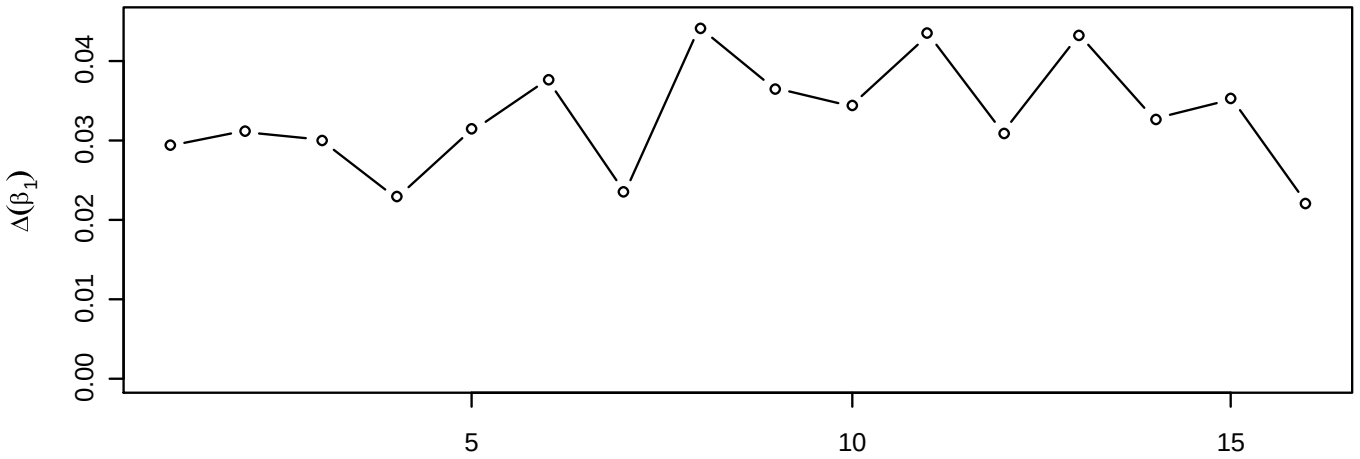
Monte Carlo Cut Offs

- x-Axis: Itemnr
- expected r^2 differences in simulated data for every item
  - <.035 negligible
  - <.07 moderate
  - >.07 large



Monte Carlo Cut Offs

- x-Axis: Itemnr
- proportional change in β between models 1 and 2 (5% is a lower baseline for DIF)



UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

SOME FINAL WORDS...

Reminder: Purposes of DIF studies

- *Purpose 1: Fairness and equity in testing.*
- *Purpose 2: Dealing with a possible threat to internal validity.*
  - rule out measurement artifact as an explanation for the group differences
- *Purpose 3: Investigate the comparability of translated and/or adapted measures.*
- *Purpose 4: Trying to understand item response processes.*
- *Purpose 5: Investigating lack of invariance.*

Zumbo, B. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going.

Testing DIF

- DIF is important because:
  - it is a way to address the unidimensionality assumption of IRT models
  - it ensures measurement invariance with respect to the criteria available and heightens test fairness

Testing DIF

- Consider both tests with and without IRT assumptions
- MH well established; logistic regression usually easy to handle
- big sample sizes
- consider corrections for multiple testing
- significance AND effect size show DIF effect
- when "cleaning" the scale of DIF items: afterwards cross-validation necessary!