UNIVERSITY OF CAMBRIDGE

The Psychometrics Centre

Summer School in Applied Psychometric Principles

Peterhouse College

13th - 17th September 2010

Introducing concepts of measurement invariance. Investigating Differential Item Functioning (DIF) using various approaches (Mantel-Haenszel and Confirmatory Factor Analysis (CFA) with covariates).

Day 4

Anna Brown, PhD
University of Cambridge

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

IRT models have desirable features:

- Items and Examinees on the same scale
  - Especially helpful in test design and score reporting
- Person parameter invariance
  - *Examinee Parameters are independent of the particular test items* - critical in computer adaptive testing and randomised testing.
- Item parameter invariance
  - *Item Parameters are independent of the examinees used to calibrate them [within a linear transformation]* - useful in field-testing and item banking.

Person parameter invariance

- Examinee’s proficiency estimates should be the same regardless which items they took
  - Impossible in CTT, but in IRT items are placed on the latent trait continuum
- Consider a test with calibrated items (item parameters have been established)
- We can randomly split items in the test and estimate the examinees’ ability from either set
  - Both estimates will be very similar
  - However, precision of estimates might differ
    - Precision will be higher when items’ locations are closer to the person parameter

Indeterminacy of the latent trait scale

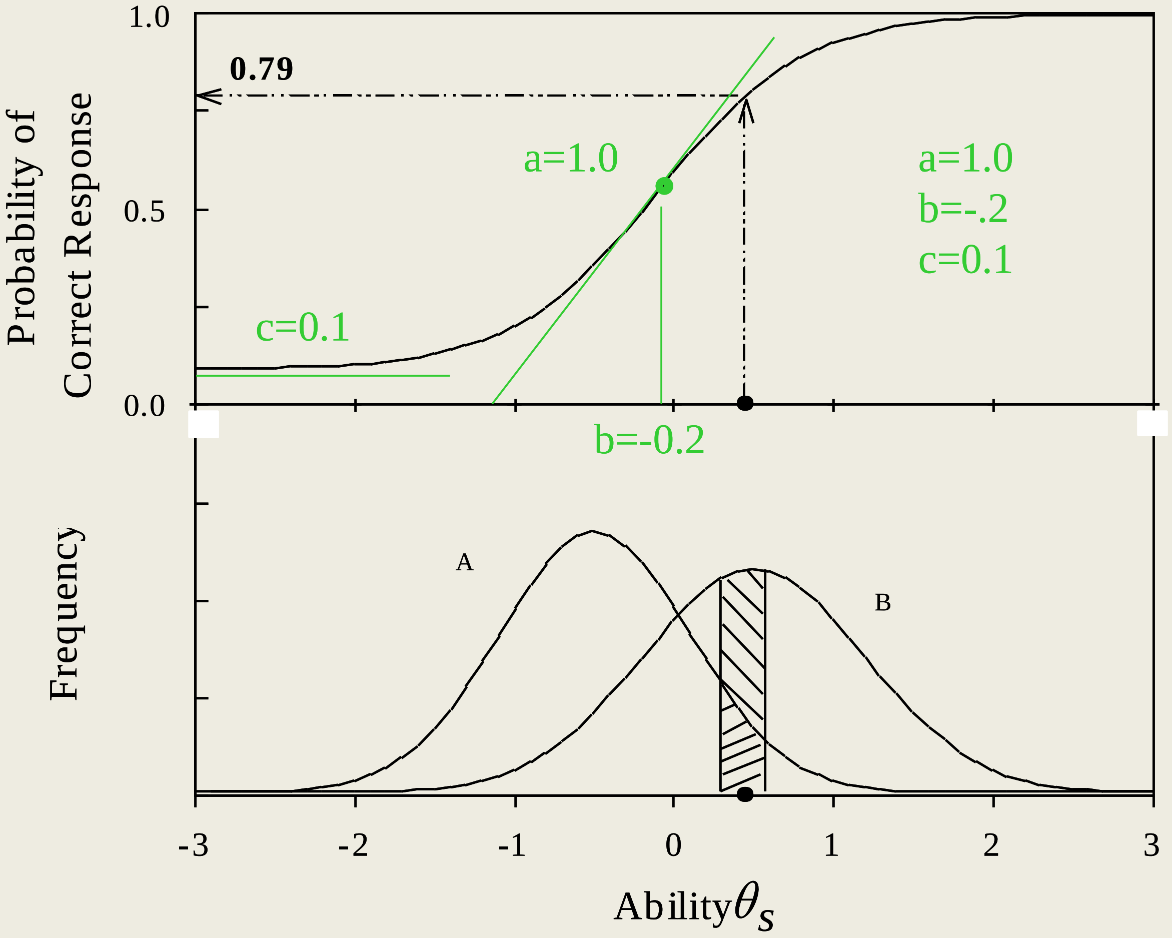
- The latent trait has no scale of its own
  - In 1PL, 2PL and 3PL models, the latent trait is assumed to have the mean=0 and SD=1 to scale the item parameters
  - In Rasch model, discrimination parameter is set to 1 and the sum of difficulty parameters is set to 0 to set the scale
- Item difficulty and discrimination parameters depend on the calibration sample properties (mean and SD)
  - In a high ability sample, item difficulties will be estimated as low
  - In a low ability sample, item difficulties will be estimated as high

Item parameter invariance

- It can be shown that difficulty and discrimination parameters in IRT are sample-invariant within a linear transformation
  - No matter who you administer the test to, you should get item parameters that relate to each other linearly
  - This is a massive advantage over CTT, where no relationship exists between item properties across different groups
- However, precision of estimates will differ
  - If there is little variance in a sample, the item will have unstable parameter estimates
- The property of item parameter invariance is very important in equating and linking
- Assessment of Differential Item Functioning rests on this property

ITEM IMPACT AND ITEM BIAS

Probability of correct answer in different ability groups



UNIVERSITY OF CAMBRIDGE

The Psychometrics Centre

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 trait
  - 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
- Analyses of item impact and DIF are *statistical* in nature

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
- Analyses of item bias are *qualitative*
- 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?

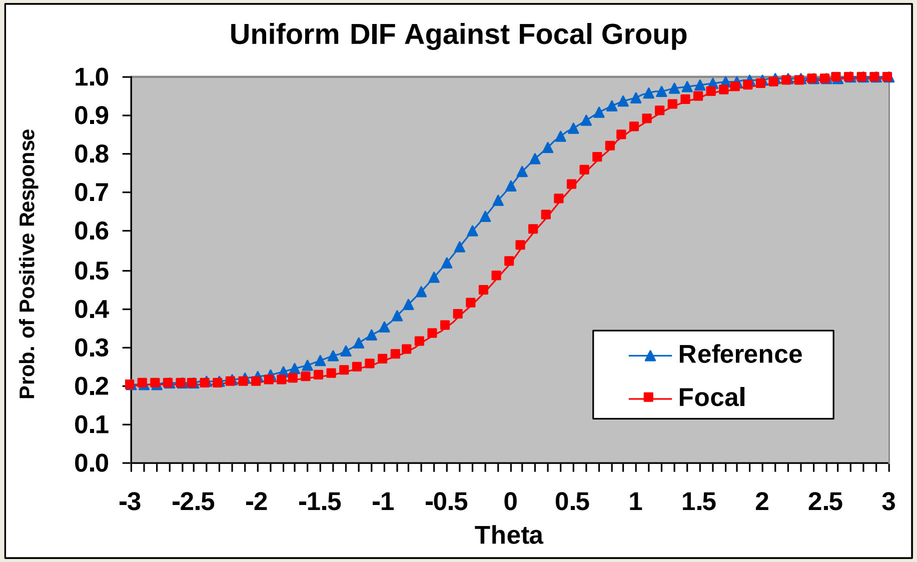
- Example 1. Students are asked to compare the weights of several objects, including a football.
  - Since girls are less likely to have handled a football, they found the item more difficult than boys, even though they have mastered the concept measured by the item (Scheuneman, 1982a).
- Example 2. 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

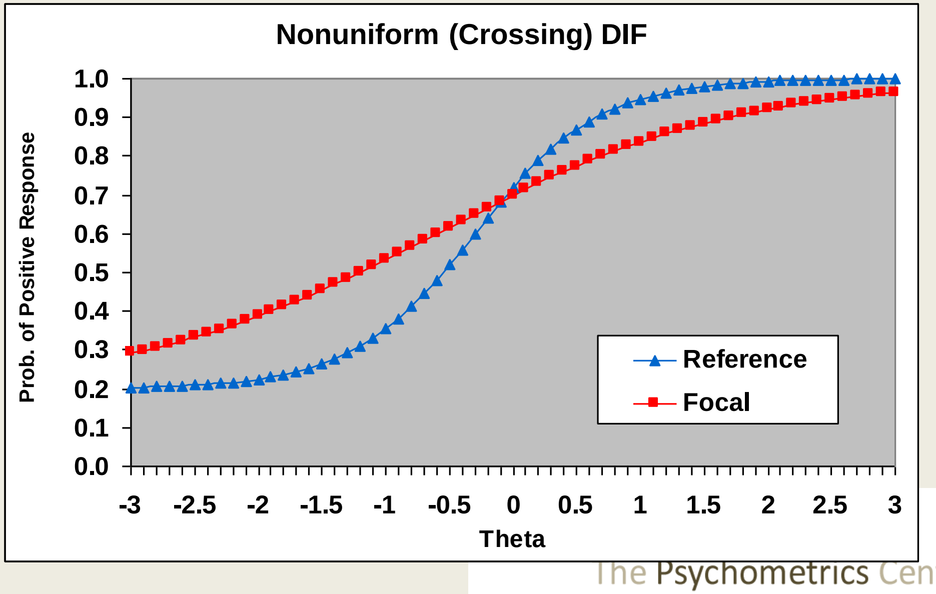
Terminology

- 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 trait
  - It can be operationalised 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 trait 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 (easy to use)
  - TestGraf (non-parametric IRT; Ramsay 1994) (difficult to use)
  - Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993)
- Parametric
  - Logistic regression (relatively easy to use but very time consuming)
  - Item Response Theory (difficult to use, but the Rasch model is easy to use)
  - Structural Equation Modelling (relatively difficult to use but extremely flexible)

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 trait

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 idea of MH method:
  - 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 1 to 10, there will be ten 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	$N_{fj} = c_j + d_j$
	$N_{1j} = a_j + c_j$	$N_{0j} = b_j + d_j$	$N_{\mathrm{j}} = a_j + b_j + c_j + d_j$

Mantel-Haenszel statistic

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

- Where $E\left(a_j\right) = \frac{N_{Rj} N_{1j}}{N_j}$ $\operatorname{var}\left(a_j\right) = \frac{N_{Rj} N_{1j} N_{Fj} N_{0j}}{N_j^2\left(N_j - 1\right)}$
- 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*

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

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

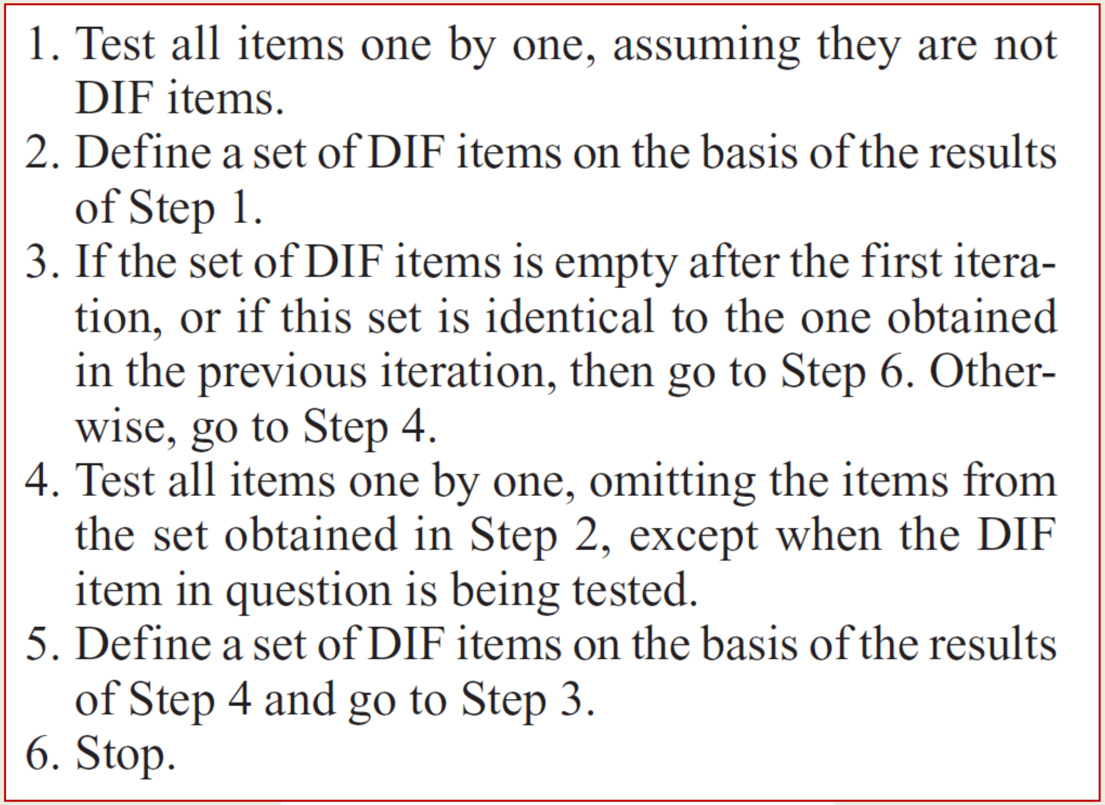
$$\lambda_{MH} = \log\left(\hat{\alpha}_{MH}\right)$$

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)

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

- Only items without DIF are used for stratification
- Item purification algorithm



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.

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

PRACTICAL – MH METHOD WITH DIFAS

DIFAS package and Practical

- *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
  - Introducing DIFAS at a workshop in Pretoria, 2008
  - Providing an example for our practical

Synthetic data generated to demonstrate DIF with dichotomous items

- Dataset courtesy Deon De Bruin, University of Johannesburg
- Synthetic data for a 15-item test with 2000 respondents
  - Respondents come from two groups (1000 per group)
- 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.

Descriptive statistics for the scale

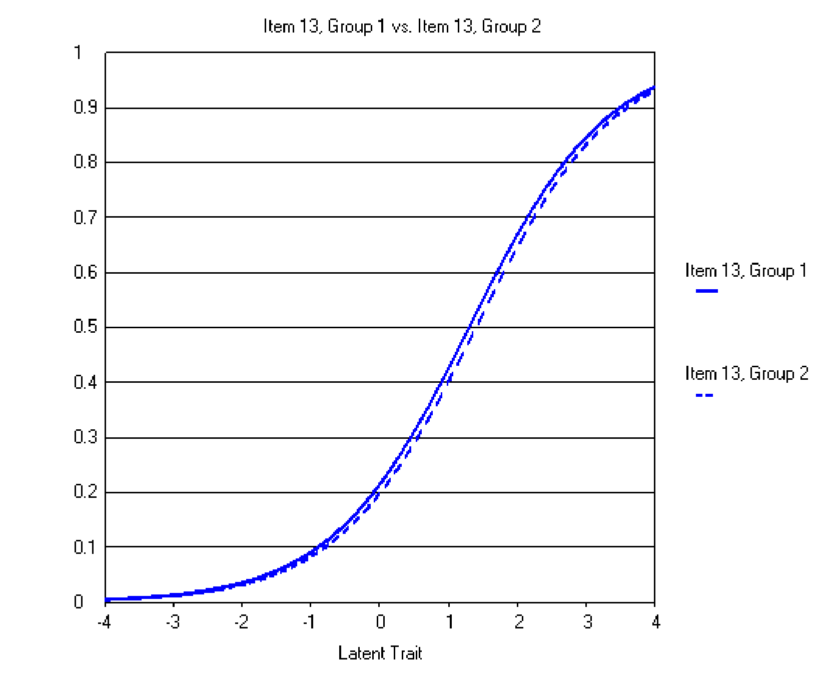
Group	Mean	SD	Cronbach alpha KR-20
Group 1 (n = 1000)	8.17	7.77	.70
Group 2 (n = 1000)	7.87	7.42	.68
Total (n = 2000)	8.02	7.61	.69

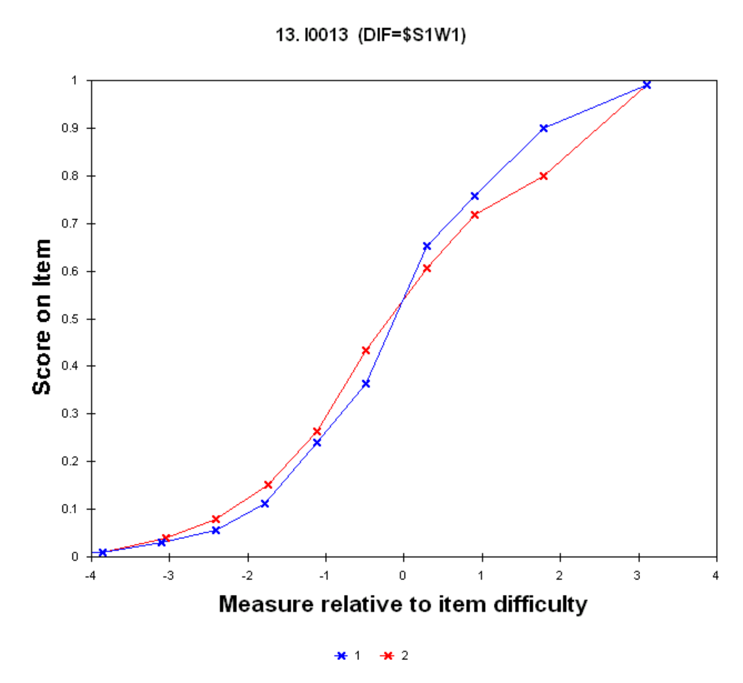
Casual inspection shows similar means, SD's and reliabilities.

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

Theoretical and empirical IRFs

- Item 13 is designed to show no DIF



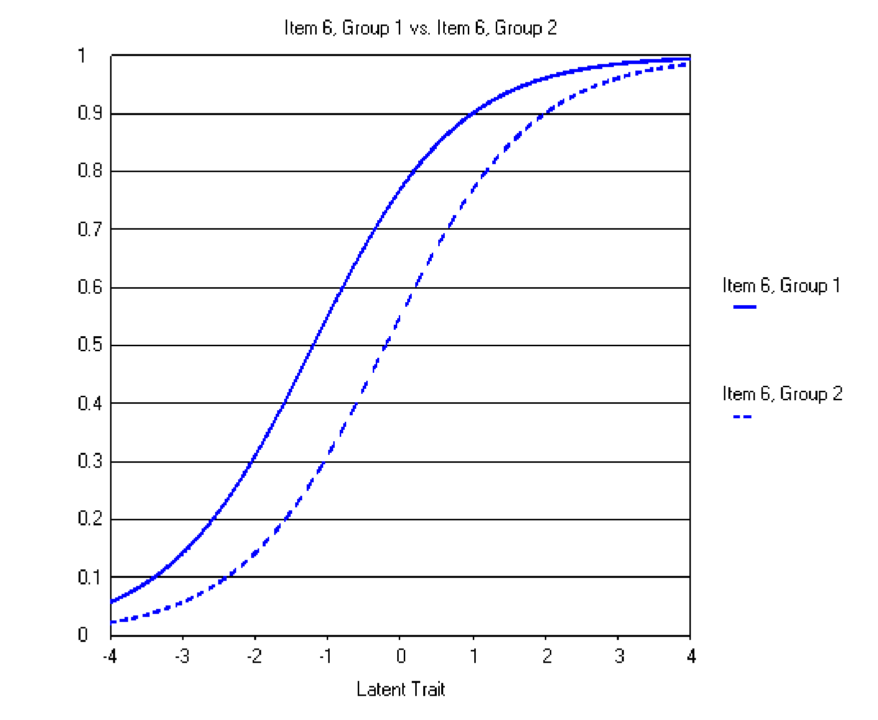


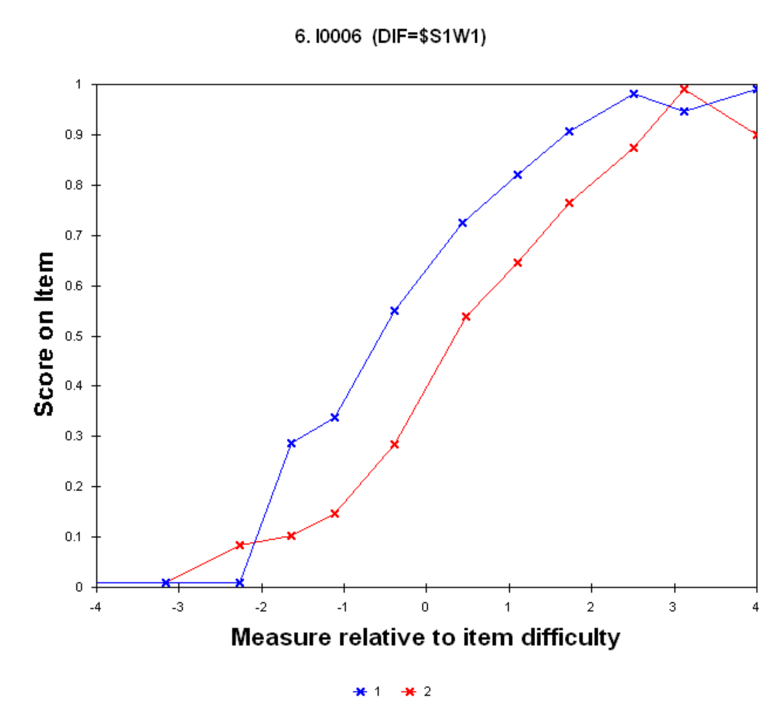
Source: De Bruin, D. (2008). What do you mean your test is cross-culturally valid? Workshop presented at SIOPSA, Pretoria, SA.

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.

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.

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

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.

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.

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.

Non-parametric DIF technique

ORDINAL MANTEL-HAENSZEL

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

Big 5 - IPIP data

- The same data we used in day 2 from IPIP (International Personality Item Pool)
- 5 symmetrical rating options:

  Very Inaccurate / Moderately Inaccurate / Neither Accurate Nor Inaccurate / Moderately Accurate / Very Accurate

- Volunteer sample, N=438 (52% female, 48% male)
- File “Big5recoded.txt” is organised as follows:
  - Gender (1=female, 2=male)
  - v1-v12 (N) v13-v24 (E) v25-v36 (O) v37-v48 (A) v49-v60 (C)

Extraversion scale

- 12 items; 4 negatively keyed items have to be recoded
- We will check items for gender DIF using DIFAS 5.0

Var	Item	Key	
14	I start conversations	1	
15	I am the life of the party	1	
16	I feel at ease with people	1	
17	I am quiet around strangers	-1	recode
18	I keep in the background	-1	recode
19	I don't talk a lot	-1	recode
20	I talk to a lot of different people at parties	1	
21	I feel comfortable around people	1	
22	I find it difficult to approach others	-1	recode
23	I make friends easily	1	
24	I don't mind being the centre of attention	1	
25	I am skilled in handling social situations	1	

Results of the Liu-Agresti estimator of the cumulative common odds ratio

```
DIF STATISTICS: POLYTOMOUS ITEMS
--------------------------------------------------------------
Name          Mantel   L-A LOR    LOR SE     LOR Z
--------------------------------------------------------------
Var 14        8.418    -0.583     0.206     -2.83
Var 15        0.319     0.107     0.189      0.566
Var 16        0.154     0.085     0.215      0.395
Var 17        0.005     0.013     0.192      0.068
Var 18        0.367    -0.118     0.195     -0.605
Var 19       14.052     -0.74     0.196     -3.776
Var 20        0.009     0.018     0.198      0.091
Var 21        0.236     0.109     0.222      0.491
Var 22         0.02    0.029     0.206     0.141
Var 23        2.093     0.304     0.206      1.476
Var 24       10.911     0.631     0.191      3.304
Var 25        0.181    -0.089     0.209     -0.426
----------------------------------------------------------
Number of strata = 15
```

3 items with statistically significant DIF (*p* < .05) were identified (printed in red). One DIF effect appears to be large.

A negative sign shows item is easier for focal group

LOR > 0.64 moderate to large DIF
LOR > 0.43 slight to moderate DIF
LOR < 0.43 slight DIF

Item bias?

- Consider item "*I don't talk a lot* "
  - Recoded, it would mean "*I talk a lot* "
  - This item is easier for the focal (female) group
    - More socially accepted from a female?
- Now consider item "*I don't mind being the centre of attention*"
  - This item is more "diffucult" for the focal (female) group
    - Less socially accepted from a female?

Differential Test Functioning

- Considering that two DIF items were favouring different groups, assessing DTF becomes important

```
DTF STATISTICS: POLYTOMOUS ITEMS
------------------------------------------------------------
Statistic                  Value          SE            Z
------------------------------------------------------------
Nu^2                       0.078        0.049        1.592
Weighted Nu^2              0.088        0.052        1.692
------------------------------------------------------------
```

With all items included the variance estimator of DTF is 0.0.078. This may be classified as small-medium DTF (close to Nu^2 threshold of 0.07). It appears that the scale functions equivalently for males and females.

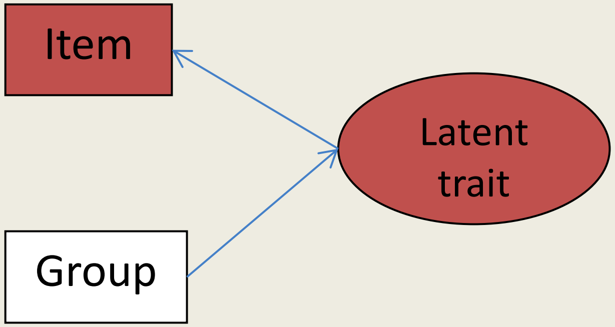
Parametric DIF technique

CFA WITH COVARIATES

Alternative way of defining DIF

- An item is unbiased if...

  item response only depends on the latent trait (i.e. group membership can only influence the item response through the trait)



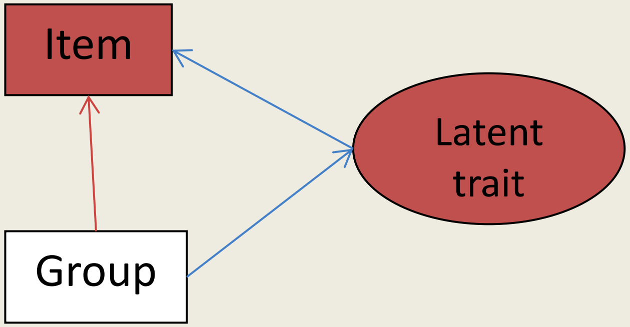
$$P(u = 1 | G, \theta) = P(u = 1 | \theta)$$

Mellenbergh, 1989

Alternative way of defining DIF

- An item shows uniform DIF if...

  item response depends on the latent trait AND the group membership (i.e. group membership influences the item response directly)



$$P(u=1|\,G,\theta)\neq P(u=1|\,\theta)$$

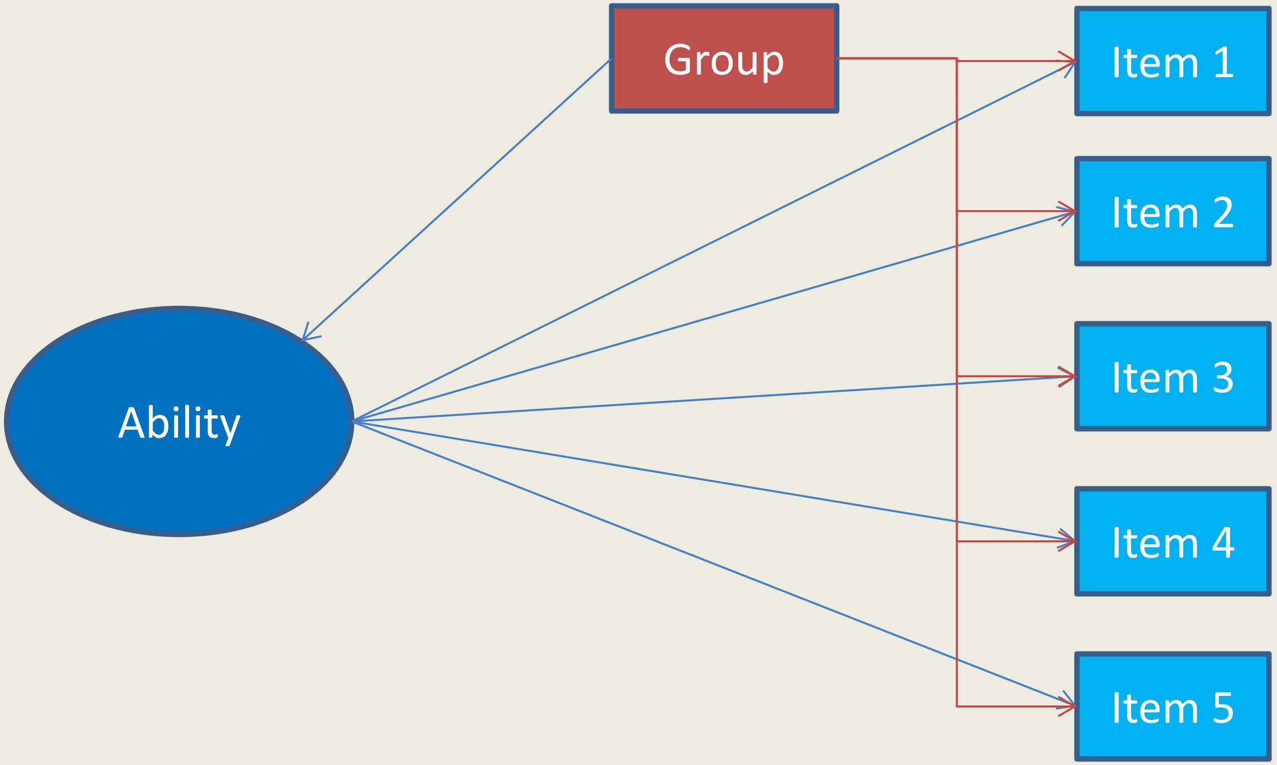
Stages of identifying uniform DIF using the covariate approach

1. Run CFA model without covariates
2. Add covariate (group indicator) but no direct effects
3. Add paths from covariate to indicators constrained to 0 - i.e. assuming there is no direct effect (Y1 on SEX@0)
4. Check modification indices
5. Add direct path from covariate to indicator for indicator with highest modification indices - rerun model
6. Repeat steps 4 & 5 until there are no further significant modification indices , evaluate model fit and significance of the direct effects

How to interpret Modification Indices

- Modification index (M.I.) is the value by which chi-square will drop if the parameter currently fixed or constrained was freely estimated
- E.P.C. is expected parameter change index
  - Expected value of the parameter if it was freely estimated

The Model



Latent Variable

Covariate

Observed items

Practical – simulated ability data

- The same data set we tested with DIFAS
- Binary ability data (15 items, 2 groups, N=1000 in each)
- DIFAS with MH non-parametric method identified several DIF items
  - Items 4,6, 12, 15 (large or C level DIF)
  - Item 10 (medium or B level DIF)
- Let’s test the same data with Mplus

Mplus syntax for Ability DIF detection

```
TITLE: testing Simulated Ability data for DIF
DATA: FILE IS dichotomousdif.txt;
VARIABLE: NAMES ARE i1-i15 group;
USEVARIABLES ARE i1-i15 group;
CATEGORICAL ARE i1-i15;
ANALYSIS:
ESTIMATOR=WLSMV;

MODEL:
ability BY i1-i15*;
ability@1;
ability ON group;
!I6 ON group*; !adding one direct effect
!i10 ON group*; !adding second direct effect

OUTPUT:  MODINDICES (ALL);
```

Mplus fit indices for Ability DIF models

Stage	Chi-square	CFI	RMSEA
1 (no direct effect)	316.533 (df=104)	0.938	0.032
2 (one direct effect)	236.556 (df=104)	0.961	0.025
3 (2 direct effects)	194.071 (df=102)	0.973	0.021
etc.			

- Please note that for all estimators apart from ML, Mplus does not allow conventional test for difference of chi-squares
- But if we use ML with categorical outcomes, we cannot request Modification indices ☹

Ability DIF – interesting results

- Stage 1 – model with no direct effects
  - Estimates a significant effect of group membership on ability

```
ABILITY  ON   GROUP       -0.139    (0.054)
```

- Stage 2 – model with one direct effect
  - Estimates NO significant effect of group membership on ability

```
ABILITY  ON   GROUP       -0.058     (0.054)
```

Parametric DIF technique

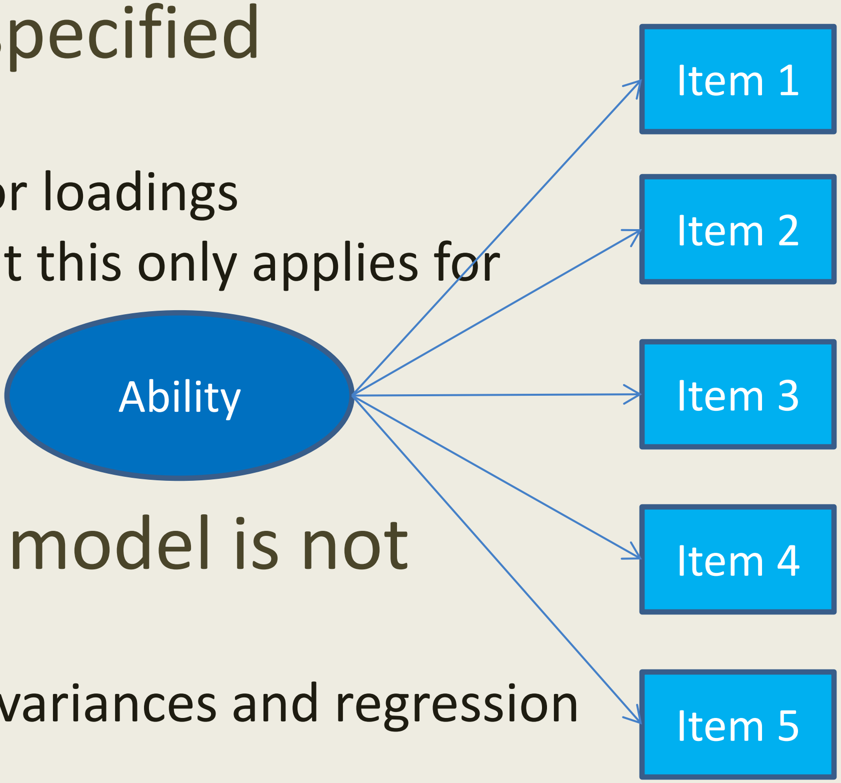
CFA WITH MULTIPLE GROUPS

CFA – multigroup approach

- Approach with covariates was only able to detect uniform DIF
- Confirmatory approach with multiple groups can be used to test for any combinations of the following
  - Measurement parameters (measurement invariance)
    - Intercepts (*item difficulty – uniform DIF*)
    - Factor loadings paths (*item discrimination – non-uniform DIF*)
    - Residual variances
  - Structural parameters (population heterogeneity)
    - Latent means
    - Latent variances/covariances/regression paths
- One of the most attractive features is that more than 2 groups can be tested

Defaults for multi-group setup

- The measurement part of the model is assumed invariant if not specified otherwise
  - Intercepts, thresholds, factor loadings
  - (except error variances – but this only applies for continuous indicators)



- The structural part of the model is not assumed invariant
  - Factor means, variances, covariances and regression coefficients

Practical – simulated ability data

- The same data set we tested with DIFAS, and with Mplus using the CFA with covariates approach
- Binary ability data (15 items, 2 groups, N=1000 in each)
- There is uniform DIF for some items, but no non-uniform DIF
- We have identified several DIF items
  - Items 4,6, 12, 15 (large or C level DIF)
  - Item 10 (medium or B level DIF)
- Let's test the same data with the multigroup approach

Mplus syntax for multi-group CFA

```
TITLE: testing equivalence of Simulated Ability data
DATA: FILE IS dichotomousdif.txt;
VARIABLE: NAMES ARE i1-i15 group;
USEVARIABLES ARE i1-i15 group;
CATEGORICAL ARE i1-i15;
GROUPING IS group (1=reference, 2=focal);
ANALYSIS:    ESTIMATOR=WLSMV;
MODEL:
     ability BY i1-i15*;
     ability@1;

!MODEL focal: [i6$1*]; ! Relax equality constraint on threshold of item 6

OUTPUT:  MOD (10);
```

Examining modification indices

```
Means/Intercepts/Thresholds
                         M.I.      E.P.C
Group=REFERENCE
[ I4$1    ]            18.753     0.075
[ I6$1    ]            60.426    -0.155
[ I10$1   ]             26.613   -0.123
[ I12$1   ]             35.066    0.132
Group=FOCAL
[ I4$1    ]            18.766    -0.412
[ I6$1    ]            60.370     0.509
[ I10$1   ]             26.585    0.211
[ I12$1   ]             35.069   -0.330
```

Note big differences in expected parameter estimates (E.P.C.) between Reference and Focal groups

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Fit indices for multi-group models

Condition	Chi-square	CFI	RMSEA
Models are equal	320.415 (df=194)	0.961	0.026
Threshold for i6 unequal	264.723 (df=193)	0.978	0.019
Thresholds for i6 and i10 unequal	225.593 (df=192)	0.990	0.012
Thresholds for i6, i10 and i12 unequal	206.835 (df=191) n/s	0.995	0.009

- Where do we stop?
- Statistical or practical significance?

Practical (Big5 IPIP- Extraversion)

- Tested before with Mantel-Haenszel
  - Found that v19 and v24 displayed uniform DIF
- Now test with the multi-group approach in Mplus

Var	Item	Key
14	I start conversations	1
15	I am the life of the party	1
16	I feel at ease with people	1
17	I am quiet around strangers	-1
18	I keep in the background	-1
19	I don't talk a lot	-1
20	I talk to a lot of different people at parties	1
21	I feel comfortable around people	1
22	I find it difficult to approach others	-1
23	I make friends easily	1
24	I don't mind being the centre of attention	1
25	I am skilled in handling social situations	1

Mplus syntax for multi-group Extraversion model

```
TITLE: GRM model testing Extraversion data
DATA: FILE IS Big5.dat;
VARIABLE: NAMES ARE gender v2-v61;
    USEVARIABLES ARE gender v14-v25;
    CATEGORICAL ARE v14-v25;
    GROUPING IS gender (1=reference, 2=focal);
ANALYSIS:       ESTIMATOR=WLSMV;  PARAMETERIZATION=THETA;
MODEL:    !Graded Response Model
Extra BY v14-v25*;
Extra@1;
!MODEL focal: [v19*];   !de-comment these later
!                   [v24*];
OUTPUT:  MODINDICES(ALL);
```

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Extraversion test results

- Fit for constrained model

```
Chi-Square Test of Model Fit          283.450*
      Degrees of Freedom              155
```

- Modification indices

```
Group REFERENCE
WITH Statements

V21    WITH V16       31.280   0.575    0.575     0.713
V24    WITH V15       11.008   0.214    0.214     0.297
```

Here is something new – in Reference group (males) we found correlated residuals

```
Means/Intercepts/Thresholds

[ V19    ]        11.746   0.397    0.397     0.272
[ V24    ]        12.043   0.298    0.298     0.299
```

Here are the items we spotted before as uniform DIF

Broader picture of test bias

DIF AND MEASUREMENT INVARIANCE

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

- Purifying the matching variable
  - taking DIF items out and re-computing scale score, and DIF again
  - Item under examination should be always included in its own matching score (Holland and Thayer, 1998)
- 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

- Adjust for DIF in the model – in M*plus* can do this by adding direct effect between the covariate and the item
- 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

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

Van de Vijver & Poortinga

UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Coming in day 5…

- DIF detection techniques implemented in R
- Applications of Item Response Theory