Measurement invariance and Differential Item Functioning

Short course in Applied Psychometrics

Peterhouse College, 10-12 January 2012

This course

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

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

UNIVERSITY OF
CAMBRIDGE
The Psychometrics Centre

RDI
Researcher
Development
Initiative

Tutors

Tim Croudace, PhD tjc39@cam.ac.uk

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

University of Cambridge, Department of Psychiatry

http://www.psychiatry.cam.ac.uk

Course content

1. What is Measurement Invariance (MI)?
   - Formal definition and how it is operationalised
2. Start with investigations of MI for binary items (Item Response Theory-based detection of Differential Item Functioning)
3. Follow up with investigations of MI for ordinal items
4. Continue with investigations of MI for continuous variables
   - Here we reinforce and consolidate our understanding of MI; introduce levels and concepts of factorial invariance
5. Finish with special (more complex to model) case of invariance of repeated measures (longitudinal MI)

- Practical sessions throughout. We will use purpose-built package DIFAS, freeware for statistical computing R, and a general modelling package M*plus*.

Colour-coding of the slide titles

- White background – lecture material
- Blue background – instructions for practical sessions
- Peach background – answers for practical sessions

What is measurement invariance

BACKGROUND

Background

- Growing impact of psychometrics
  - Educational testing
  - Workplace testing
  - Clinical trial outcome evaluations
  - Health care interventions etc.
- Psychometrics is controversial
  - Adverse impact on some individuals if their test scores are *biased* in any way
  - Can lead to
    - breach of equal opportunities
    - misdiagnosis in medical practice
    - inequality of opportunity in education
    - wrong conclusions in research

Possible sources of bias

- Construct bias
  - Definition/appropriateness of constructs is different between groups
- Method bias
  - Instrument bias – instrument features not related to the construct (familiarity with stimulus material etc.)
  - Administration bias
  - Response bias
- Item bias
  - Item-related nuisance factors (e.g. item may invoke additional traits or abilities)
  - Poor translation in adapted tests

What is Measurement Invariance?

- …Some *properties* of a *measure* should be independent of the characteristics of the person being measured, apart from those characteristics that are the intended *focus* of the measure. (Millsap, 2007)

- Some elaboration is required
  1. What do we mean by ‘measure’?
  2. What do we mean by ‘properties’ of a measure?
  3. What is the intended ‘focus’ of the measure?

What we mean by 'Measure'

- We do not mean any specific test or type of test
  - MI should apply to individual test items, blocks of test items, subtests, or whole tests.
- MI should apply to various formats, such as self-ratings, or judgments made by other raters.
- No particular scale properties for the measure are assumed
  - MI should apply to discrete nominal or ordinal scores, or to continuous interval scores.

What we mean by ‘Properties’ of a measure

- We don’t expect all properties of a measure to be invariant.
  - The average score on a measure will generally vary
  - The reliability of a measure will generally vary, because variation in attribute may be different across groups of examinees.
- If a measure based on a common factor model, we do expect that the unstandardized factor loading(s) will be invariant under fairly broad conditions

The intended ‘focus’ of a measure

- *A priori* definition of the intended focus of the measure is required
  - So we can distinguish relevant and irrelevant properties of a measure
- In psychological measurement, the attributes that we are trying to measure are usually formally defined as latent variables.
  - The measure can be underlined by one (unidimensional) or several (multidimensional) latent variables

Operational definition of MI

- MI holds if and only if *the probability of an observed score, given the attribute level and the group membership, is equal to the probability of that given only the attribute level*.

Formal definition of MI

$$P(\mathbf{X}|\mathbf{W},\mathbf{V}) = P(\mathbf{X}|\mathbf{W}) \qquad (1)$$

- **X** = *observed scores on the measure*
- **W** = *intended latent variables for X*
- **V** = *other characteristics (often a scalar group identifier for demographic variables such as gender or ethnicity)*

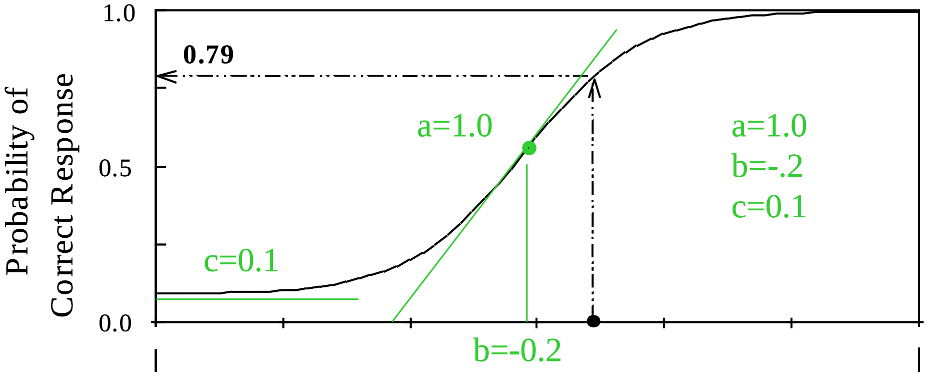
- **V** *should be irrelevant to* **X** *once* **W** *is considered*

Mellenbergh (1989), Meredith (1993)

- Depending on particular type of **X** and model for relationships between **W** and **X**, different investigations take place

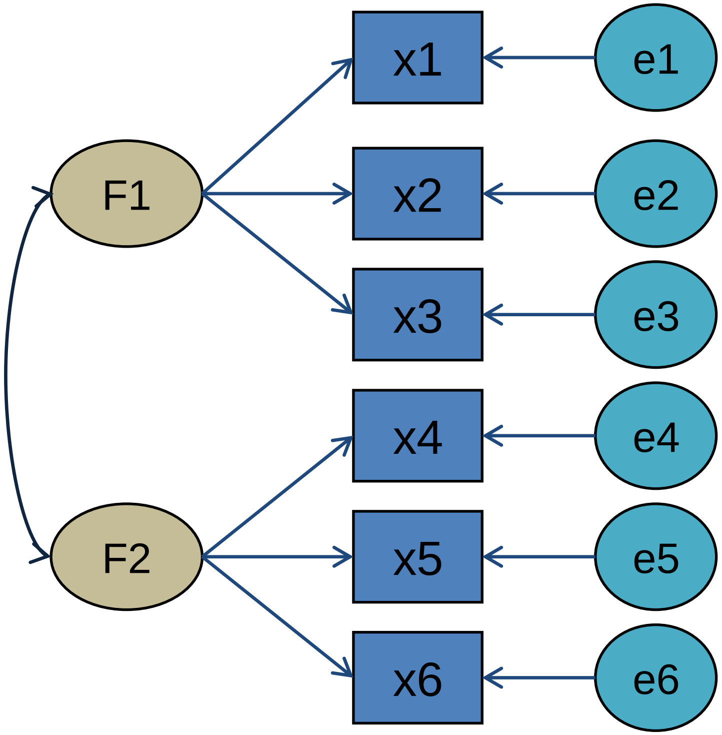
1. **X** fits an item response model

- When **X** are item scores (e.g. binary) that fit one of models in Item Response Theory (IRT),
  - **W** represents continuous latent variable(s)
  - investigations of MI evaluate Differential Item Functioning
  - DIF is directly concerned with unequal probabilities of giving a certain response on an item for members of different groups **V**, after matching on the attributes the test is intended to measure



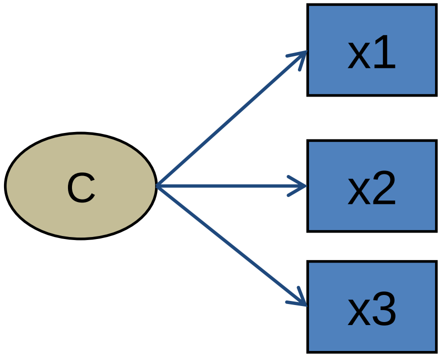
2. **X** fits a common factor model

- When **X** fit a common factor model, MI implies factorial invariance
- Factorial invariance has a long history in psychometrics
- But factorial invariance is weaker than MI in (1) because
  - only means and covariance structure (first and second moments) is studied in factorial invariance investigations
  - And (1) requires invariance in conditional distributions.



3. **X** fits a latent class model

- When **X** are item scores that fit a latent class model,
  - **W** is categorical and represents a latent class identifier
  - investigations of MI evaluate probabilities of giving a certain response on an item for members of different groups **V**, conditional on the membership in a latent class **W**
  - Beyond the scope of this course

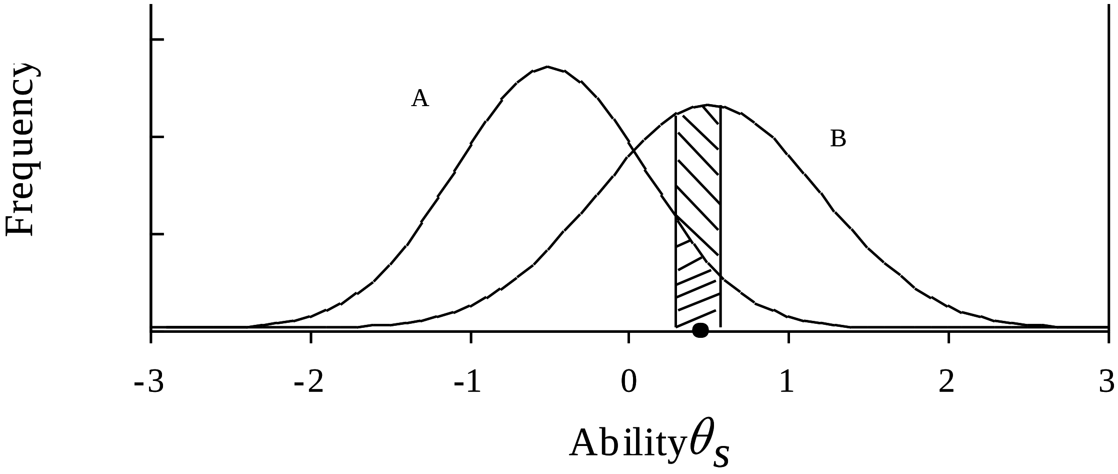


1. **X** fits an item response model

BINARY TEST ITEMS

Item impact

- 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 construct* that the item is intended to measure
- *Notice that this is exactly the definition of MI applied to test items*

Example 1

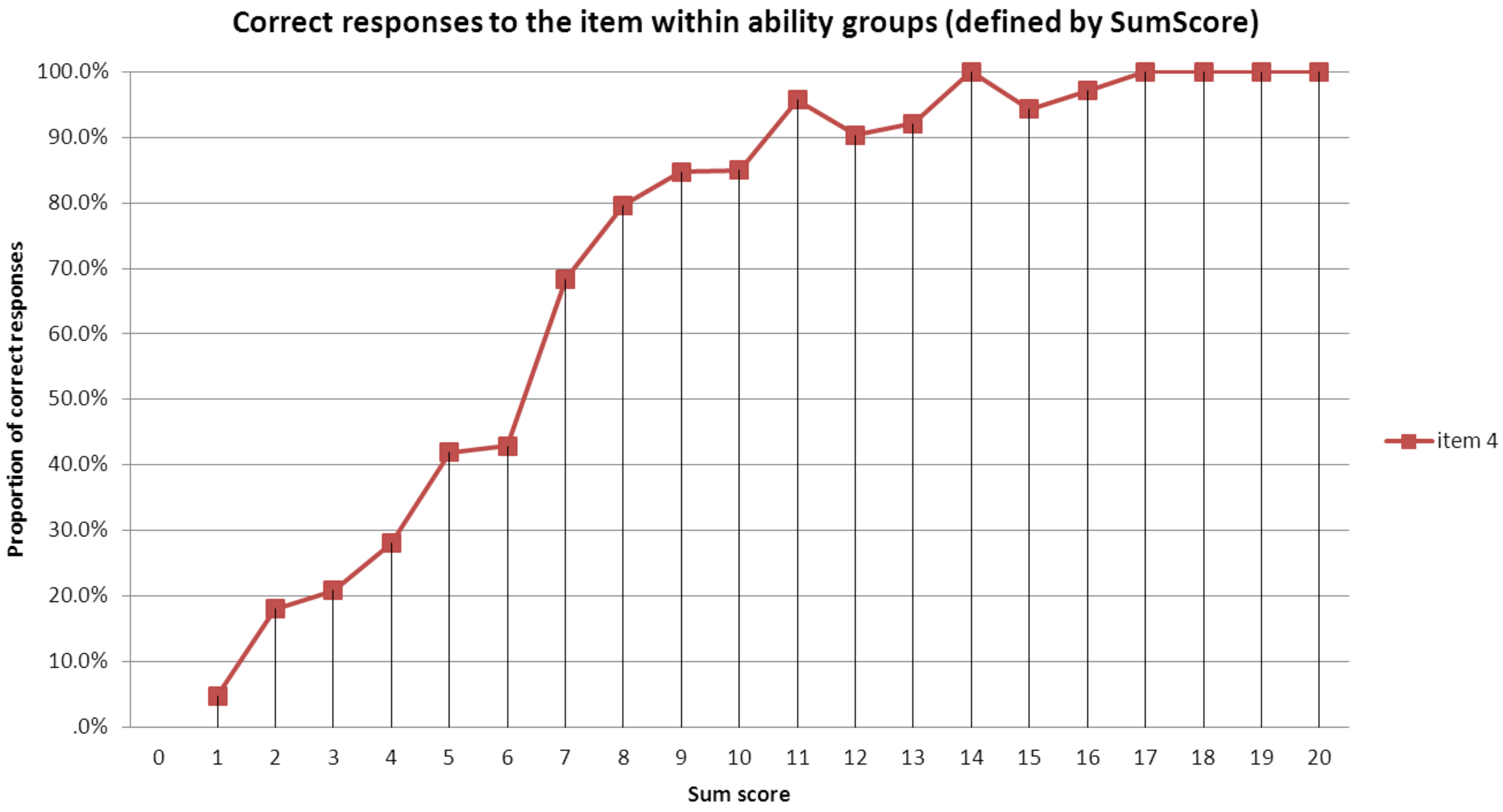
- Students are asked to compare the weights of several objects, including a football (Scheuneman, 1982).
  - Since girls are less likely to have handled a football, they found the item more difficult than boys, even though they have mastered the concept measured by the item.

Example 2

- A vocabulary test asked to find a synonym to “ebony”.
  - Ebony is a dark-coloured wood and it is also the name of a popular magazine targeted to African-Americans.
  - The Black students were more likely to answer the item correctly than the White students throughout the bulk of the test score distribution.

A STUDENT'S GUIDE TO MONEY
EBO
SUDAN: HOPE AMID DARFUR'S TRAGEDY
NEW YORK'S FIRST BLACK GOVERNOR
WHO SHOULD ADOPT OUR CHILDREN?
SPECIAL FASHION ISSUE WHERE ARE ALL THE BLACK MODELS?
PLUS FASHION YOU CAN AFFORD
HOW TO EARN $1 MILLION IN COLLEGE SCHOLARSHIPS
MAN TO MAN: HOW SEEING THE DOCTOR CAN SAVE YOUR LIFE
THE real Michelle OBAMA
THE FIRST LADY HOPEFUL
TALKS FAMILY, FRIENDSHIP & THE FUTURE OF AMERICA
INSIDE: MICHELLE WILLIAMS HERBIE HANCOCK DONNA SUMMER BISHOP T.D. JAKES AND OMAR TYREE

Likelihood of correct response as function of ability



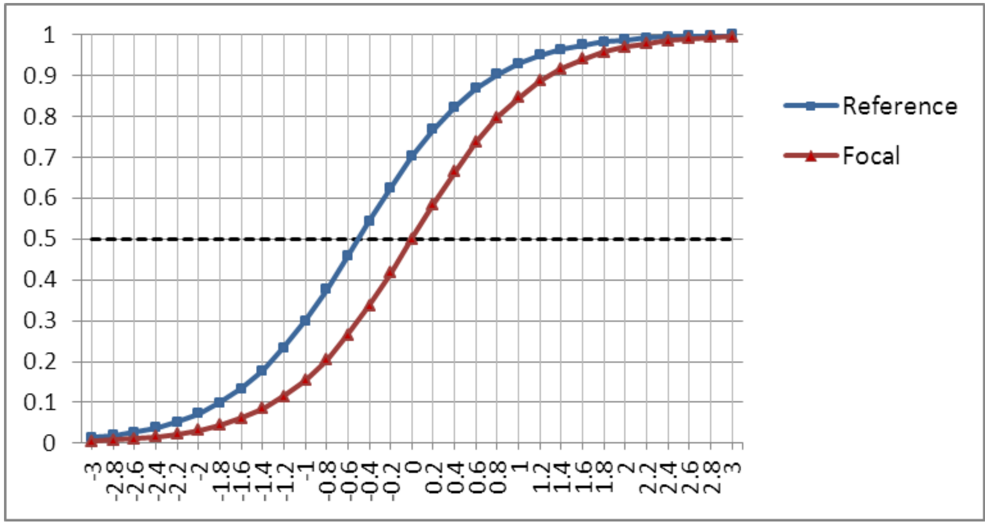
Terminology

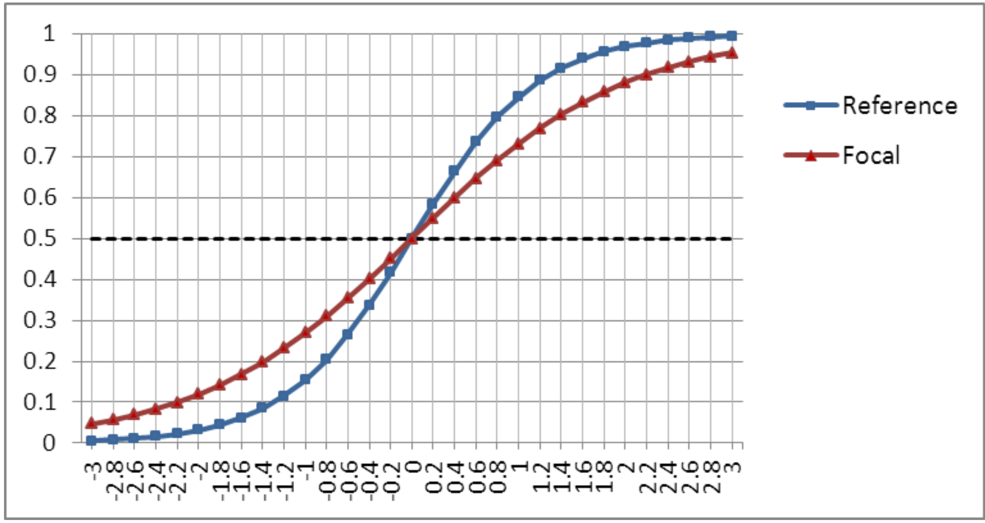
- Reference and focal groups
  - The reference group is the group that serves as the standard
  - The focal group is the group that is compared against the standard
  - Typically, the majority group or the group on which a test was standardized serves as the reference group
- Matching variable
  - Participants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

Uniform and non-uniform DIF

- Uniform DIF
  - E.g. lower probability of endorsing the item at all trait levels
  - Affects origin of scale

- Non-uniform DIF
  - Higher probability of endorsing the item at low level of trait, but lower probability at high level (or vice versa)
  - Affects measurement unit and origin of scale





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 bias & DIF

- Analyses of item bias are qualitative in nature: reconstruction of meaning and contextualization
- Analyses of DIF are 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
    - Content analysis by subject matter experts

Item bias or no item bias?

- Example 1. Students were asked to compare the weights of several objects, including a football.
  - Sheuneman argues that the item is biased against girls.
- Example 2. A vocabulary test asked to find a synonym to “ebony”.
  - The item was considered to an important part of the curriculum and was not removed from the test.

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 uniform DIF
- The MH method treats the DIF detection problem as one involving three-way contingency tables. The three dimensions of the contingency table involve
  - whether one gets an item correct or incorrect
  - group membership, while conditioning on the test score
  - the total score “sliced” into a number of category score bins.

Score “slices”

- Sum score is usually used as a matching variable
- The item being studied for DIF must be included in the sum (Zwick, 1990).
- 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

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

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.

- 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

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

ETS classification for the DIF effect size

- Educational test services (ETS) uses the following classification scheme
- ETS Delta scale (Holland &Thayer, 1988) is computed as

$$\Delta_{MH} = -2.35\lambda_{MH}$$

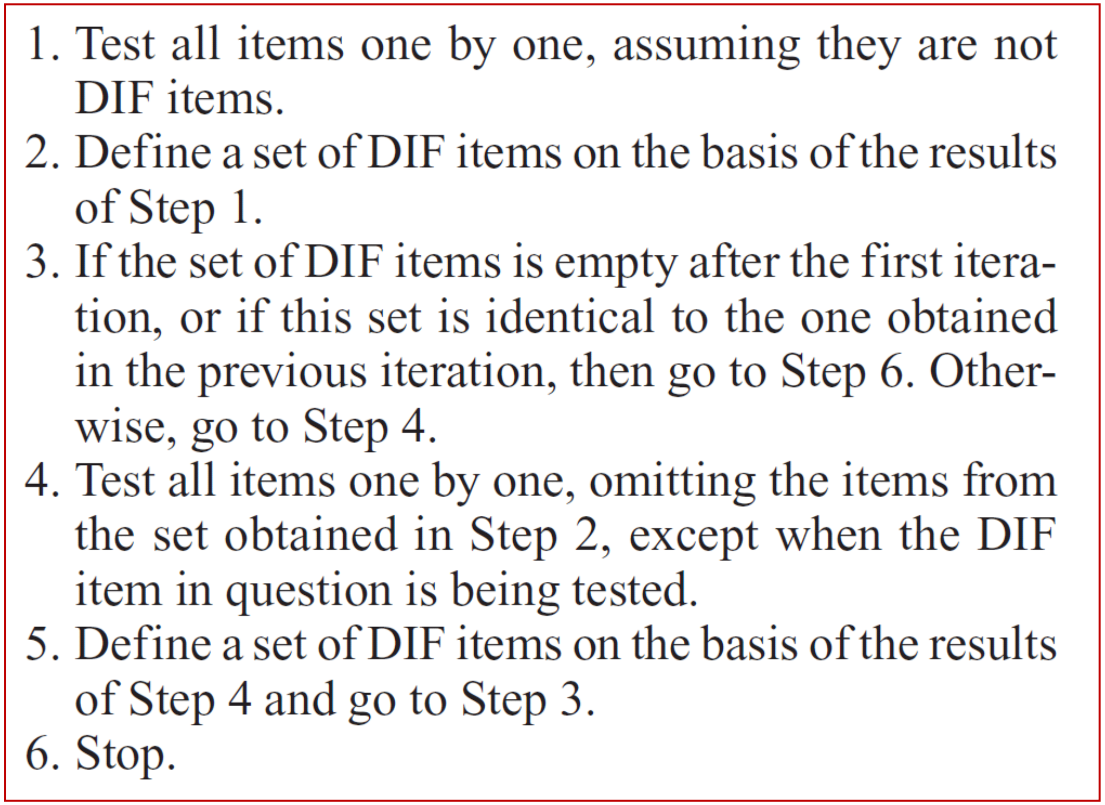
- And the following cut-offs are used
  - – Large DIF $|\Delta_{MH}|>1.5$ (Class C)
  - – Moderate DIF $1<|\Delta_{MH}|\leq 1.5$ (Class B)
  - – Small DIF $|\Delta_{MH}| \leq 1$ (Class A)

Steps in 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 DIF
  - see Penfield & Algina, 2006, p. 307.

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.

What about non-uniform DIF? Breslow-Day statistic

- Classical MH tests are only effective for uniform DIF
- A non-parameteric method for non-uniform DIF: Breslow-Day statistic
- Remember common odds ratio? (should be 1 if there is no DIF)

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

- As non-uniform DIF increases, odds-ratios become more heterogeneous, i.e. their deviation from the expected value (A) increases
- Breslow-Day statistic tests whether the odds ratios are homogeneous over the range of the scale by testing deviations from A
  - distributed approximately as chi-square with 1 degree of freedom

DIFAS package

- DIFAS covers all functions for MH-based DIF tests
  - Item purification can be done by hand
- We provide a short tutorial for DIFAS in a separate slide deck

- *DIFAS*, and its corresponding manual, can be can be downloaded free of charge from a webpage 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 at SIOPSA
  - Providing the example dataset for our Practical exercises

Illustration – NSHD Dataset

- Responses from the ongoing Medical Research Council National Survey of Health and Development (NSHD), also known as the British 1946 birth cohort.
- Wave of interviewing undertaken in 1999 when the participants were aged 53
  - *Wadsworth M.E., Butterworth, S.L., Hardy, R.J., Kuh, D.J., Richards, M., Langenberg, C., Hilder, W.S. & Connor, M. (2003). The life course prospective design: an example of benefits and problems associated with study longevity. Social Science and Medicine, 57, 2193-2205.*

- A total of N=2901 respondents (1422 men and 1479 women) provided answers to the GHQ-28.

GHQ-28 Instrument

- The 28-item version of General Health Questionnaire (GHQ-28)
  - *Goldberg, D. P. (1972). The detection of psychiatric illness by questionnaire. Oxford University Press: London.*
  - Developed as a screening questionnaire for detecting non-psychotic psychiatric disorders in community settings and non-psychiatric clinical settings
- Respondents are asked to think about their health in general and any medical complaints they have had over *the past few weeks.*

- Rating scale with 4 alternatives
  - slightly different for each item, in phrasing and verbal anchors
- Example question

"Have you recently lost much sleep over worry?"

(Not at all - No more than usual - Rather more than usual - Much more than usual)

GHQ-28 a priori structure

- Designed to measure 4 *a priory* facets of mental health variation (measured with 7 items each)
  1. Somatic Symptoms,
  2. Social Dysfunction,
  3. Anxiety / Insomnia,
  4. Severe Depression / Hopelessness.

- Also, the general *psychological distress* factor can be measured

Subscale A: somatic symptoms

Have you recently

Al been feeling perfectly well and in good health?
A2 been feeling in need of a good tonic?
A3 been feeling run down and out of sorts?
A4 felt that you are ill?
A5 been getting any pains in your head?
A6 been getting a feeling of tightness or pressure in your head?
A7 been having hot or cold spells?

Not at all	*No more than usual*	*Rather more than usual*	*Much more than usual*

Stratum-level frequencies

```
----------------------------------------------------------
                    Reference    Focal
Stratum             Frequency    Frequency
----------------------------------------------------------
0                   1076         925
1                   125          222
2                   94           87
3                   47           68
4                   47           96
5                   26           38
6                   3            27
7                   4            16
----------------------------------------------------------
```

males ***females***

Examining gender DIF

Breslow-Day, and the Combined Decision Rule in the next column

```
---------------------------------------------------------------------------
Name      MH CHI     MH LOR     LOR SE     LOR Z          BD       CDR      ETS
---------------------------------------------------------------------------
Var 1     3.5263     0.3698     0.1882     1.9649      0.369        OK        A
Var 2     7.2384     0.4454     0.1658     2.6864     10.326      Flag        B
Var 3     0.5242     0.15       0.1859     0.8069      3.532        OK        A
Var 4    14.8900     0.8101     0.2081     3.8928      0.847      Flag        C
Var 5     0.0818    -0.0946     0.2355    -0.4017      1.637        OK        A
Var 6     0.2905     0.1516     0.2307     0.6571      0.08         OK        A
Var 7    64.4739    -1.3489     0.1756    -7.6817     10.208      Flag        C
---------------------------------------------------------------------------
                                             Reference Value = 0, Focal Value = 1
```

A negative sign shows the item is 'easier' for the focal group;
A positive sign shows the item is more 'difficult'

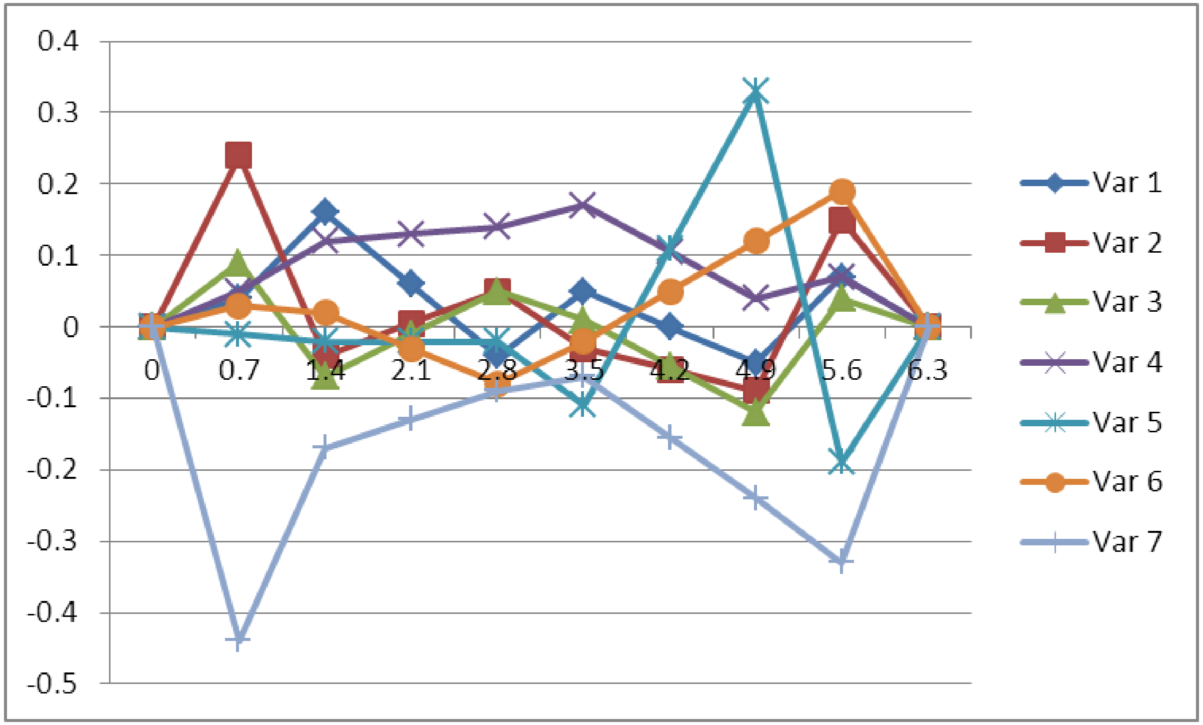
DIFAS gives a useful breakdown

- DIFAS prints differences in conditional probabilities between groups for score intervals

```
CONDITIONAL DIFFERENCES: Intervals of size 0.7
--------------------------------------------------------------------------------
Lower    0     0.7     1.4    2.1     2.8     3.5    4.2     4.9     5.6     6.3
Upper  0.7     1.4     2.1    2.8     3.5     4.2    4.9     5.6     6.3     7.1
--------------------------------------------------------------------------------
Var 1    0    0.04    0.16      .   -0.04    0.05      .   -0.05    0.07       0
Var 2    0    0.24   -0.04      .    0.05   -0.03      .   -0.09    0.15       0
Var 3    0    0.09   -0.07      .    0.05    0.01      .   -0.12    0.04       0
Var 4    0    0.05    0.12      .    0.14    0.17      .    0.04    0.07       0
Var 5    0   -0.01   -0.02      .   -0.02   -0.11      .    0.33   -0.19       0
Var 6    0    0.03    0.02      .   -0.08   -0.02      .    0.12    0.19       0
Var 7    0   -0.44   -0.17      .   -0.09   -0.07      .   -0.24   -0.33       0
--------------------------------------------------------------------------------
```

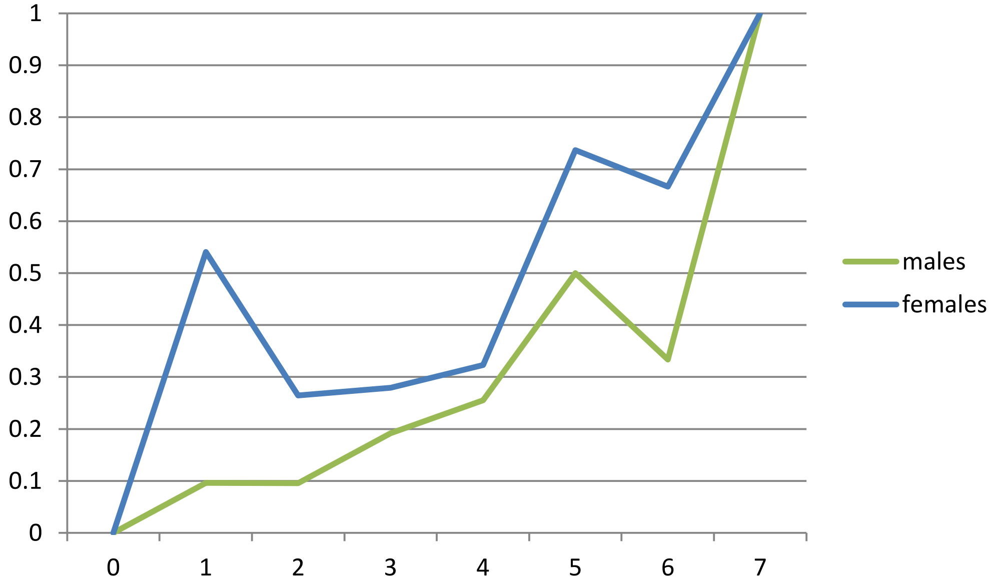
Plotting differences in conditional probabilities by score

- Notice that for item 4, differences in conditional probabilities are largest in the middle of the scale
- For item 7, differences in conditional probabilities are largest at the extremes



Empirical proportions of endorsement for item 7

- Plotting empirical proportions of endorsement for item 7 for each level of the sum score
- Discrepancies are heterogeneous
  - This explains significant Breslow-Day statistic
  - Could be interpreted as non-uniform DIF; however, DIF here is clearly uniform as conditional probabilities are always higher for females



Item bias?

- Item 7 shows large DIF in favour of the focal group (females)
- Consider item content: "*have you recently been having hot or cold spells?*"
  - This item has a higher intercept in the female group
  - Females endorse it much more easily than males with the same level of somatic symptoms
  - Consider the age of the cohort at the moment of testing (53 years)
  - Is this symptom in females indicative of general health in the same way as in males?

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

Quantifying DTF

- 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, τ^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

Gender DTF with all items included

```
----------------------------------------------------
Statistic         Value          SE             Z
----------------------------------------------------
Tau^2             0.365          0.216          1.69
Weighted Tau^2    0.476          0.274          1.737
----------------------------------------------------
```

With all items included the variance estimator of DTF is 0.365. This is large DTF (Tau^2 > 0.14).

Gender DTF with item 7 excluded

```
---------------------------------------------------
Statistic          Value         SE            Z
---------------------------------------------------
Tau^2              0.072         0.068         1.059
Weighted Tau^2     0.047         0.05          0.94
---------------------------------------------------
```

With the largest DIF item (item 7) excluded, the variance estimator of DTF is 0.072. This is just on the cut-off for small to medium DTF (Tau^2 < 0.07).

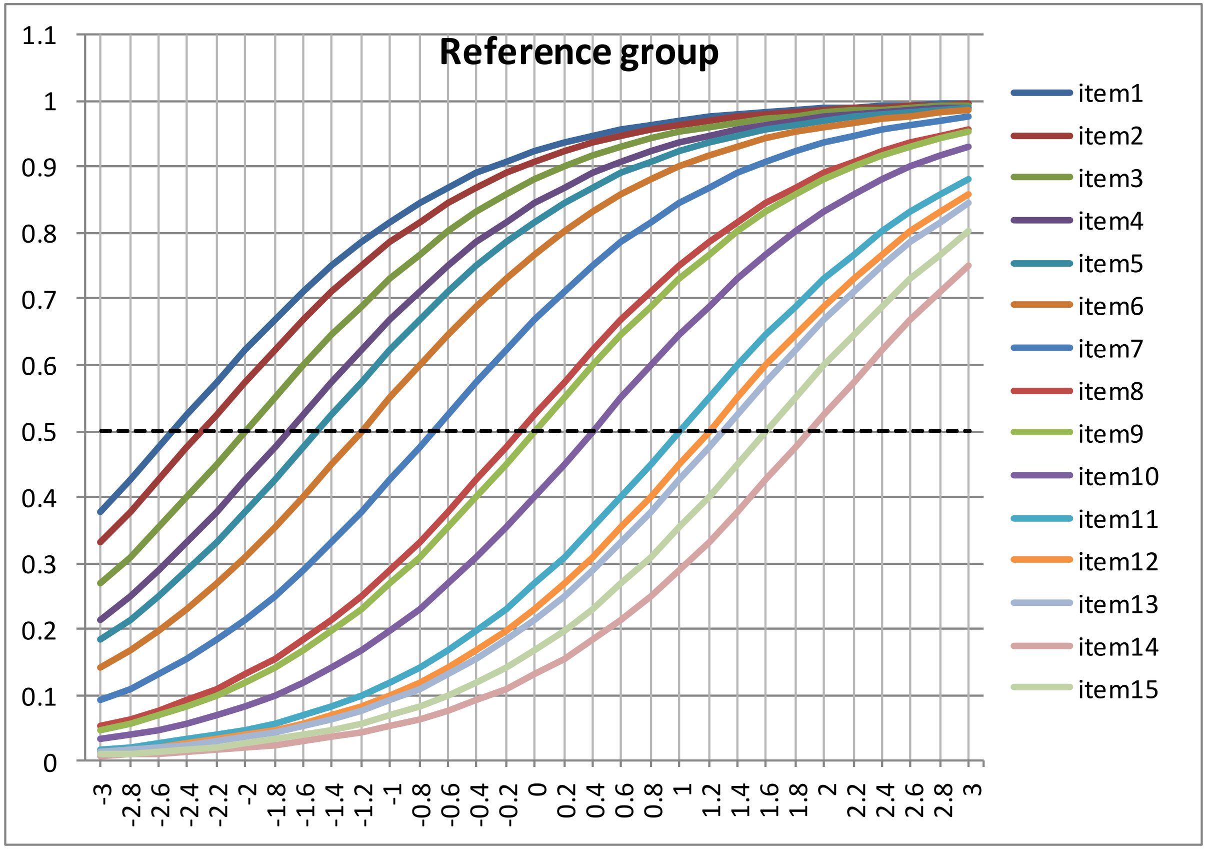
MH for dichotomous items with DIFAS

ABILITY DATA FOR PRACTICAL 1

Identifying DIF with dichotomous items

- Source:
  - De Bruin, D. (2008). *What do you mean your test is cross-culturally val*id? Workshop presented at SIOPSA, Pretoria, SA.
- Synthetic data for a 15-item test with 2000 respondents
  – Respondents come from two groups (1000 per group)
- The data were generated as follows
  – All the items have equal loadings
  – For six items the intercepts were specified to differ across groups
  – Hence, six items have uniform DIF, but no items have non-uniform DIF
  – The ability of the two groups is equal

ICCs for the Reference group



True item “difficulties” (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's alpha
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

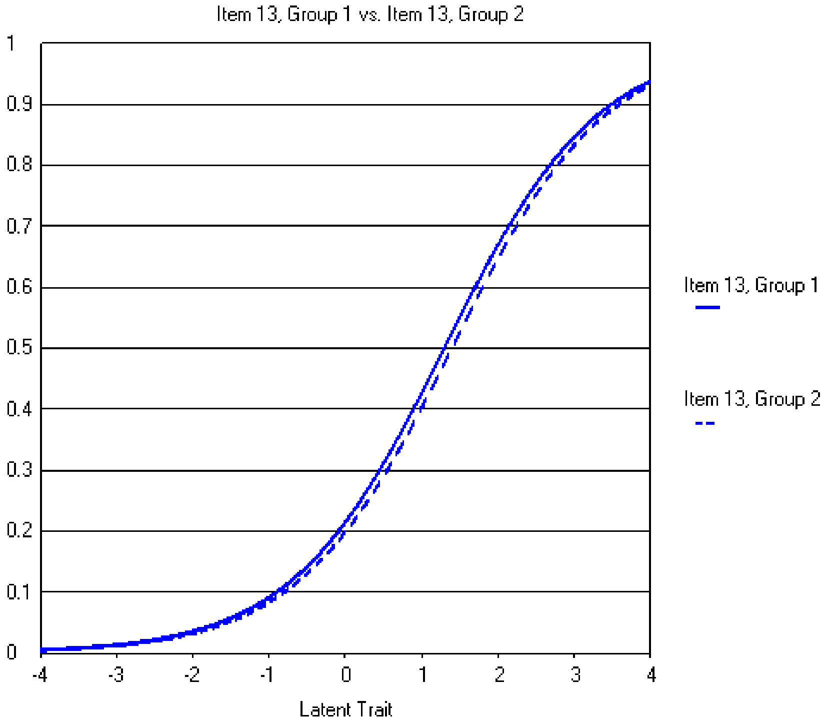
Casual inspection shows similar means, SD's and reliabilities.

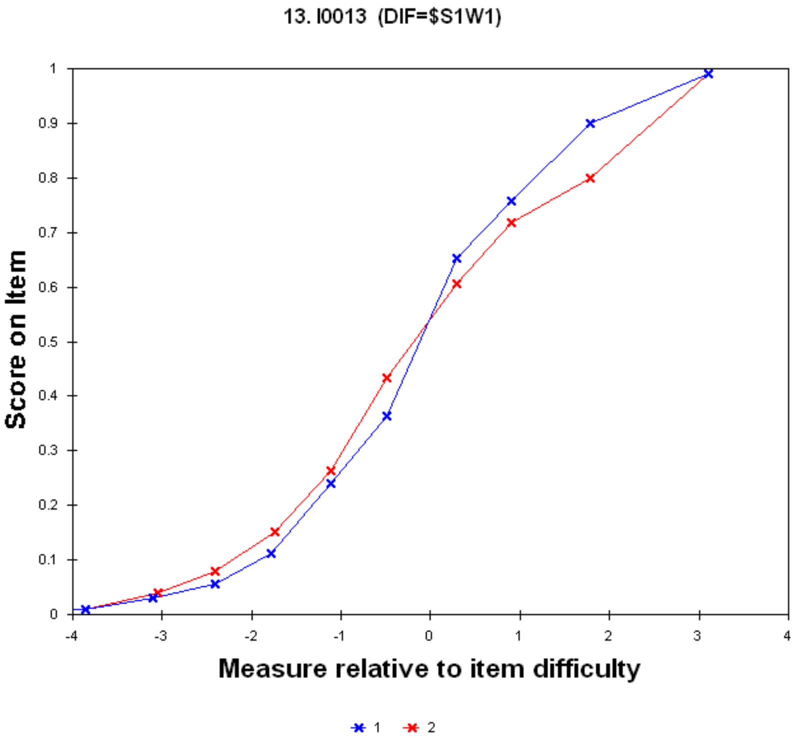
Source:

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

Theoretical and empirical ICCs

- Item 13 is designed to show no DIF



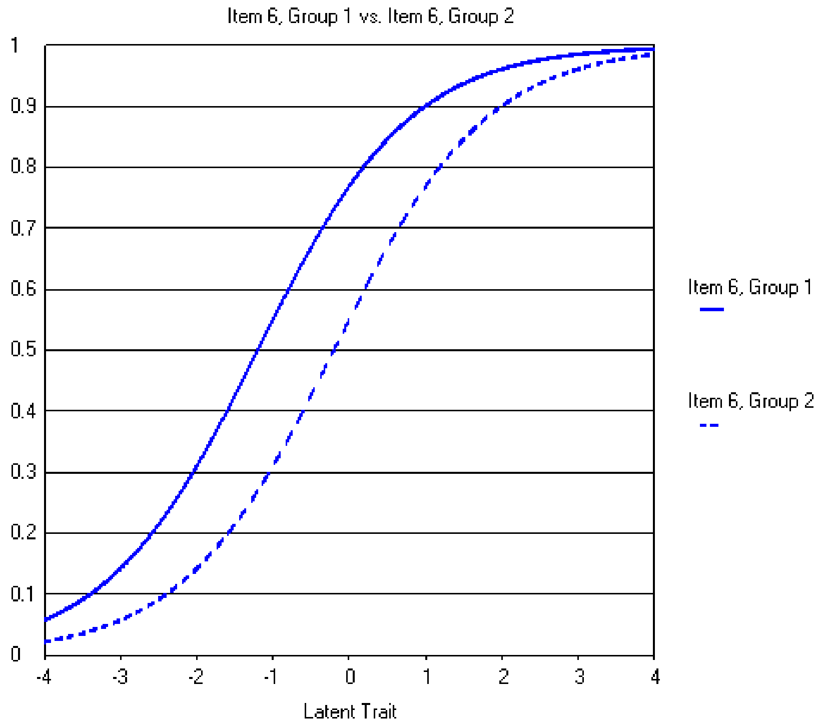


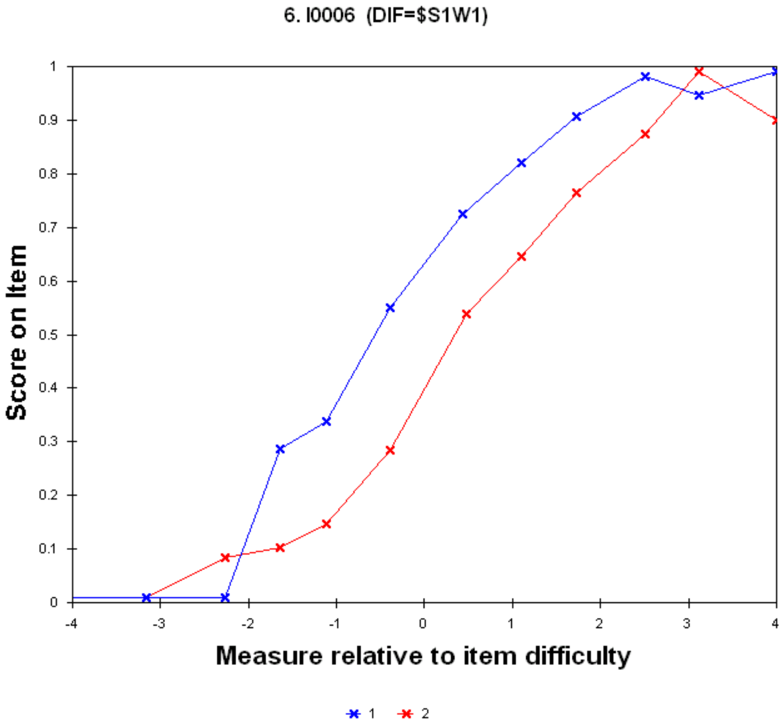
Source:

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

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.

Practical 1.
INSTRUCTIONS for MH DIF study

- Open the folder ‘Practical 1’ and find DIFAS program and manual, and data file ‘dichotomousDIF.txt’
- Start DIFAS software
- Use menu to open the data file and specify the items to be studied, the matching variable and the size of the slices
- Run DIF analysis and interpret the results

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

```
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 the item is easier for the focal group

Breslow-Day is significant here. This is a false positive due to chance draw of a sample from all simulated samples

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

Practical 1.
INSTRUCTIONS for MH DTF study

- Continue where we left off with ‘dichotomousDIF.txt’
- Run DTF analysis with all items included and interpret the results
- Exclude the worst DIF items and repeat the DTF analysis

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 is 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?

Source:

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

MANTEL-HAENSZEL WITH R

Statistical computing with R

- **R** is an open source statistical computing environment
- **R** is a console, so no "GUI" / point-&-click interface is available
- Everything is in code, e.g.: reading data:

```
GHQ28 <- read.table(file.choose(),
 header=TRUE, sep="\t", na.strings="NA",
 dec=".", strip.white=TRUE)
```

- R is case sensitive, so ghq28 and **GHQ28** are different things!

R package for DIF analysis (difR)

- **difR** is a package that provides several functions to identify *dichotomous* DIF
  - Magis, D., Béland, S., Tuerlinckx, F., & de Boeck, P. (2010)
- to load **difR**, type: `library(difR)`
- it refers to the **ltm** package which has to be also installed

difR package functions

- Mantel-Haenszel procedure

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

  - Requires an object containing the items
  - Requires a grouping variable
  - Requires a code for the Focal group
  - MHstat is either `"MHChisq"` or `"logOR"`

- Breslow-Day procedure

```
difBD(Data, group, focal.name, purify=FALSE,
   Bdstat= "trend")
```

Running Mantel-Haenszel for GHQ28 somatic symptoms: `difMH`

- Create grouping variable and item set

```
gender <- GHQ28[ ,29]
somatic <- GHQ28[ ,1:7]
```

- Call the MH function without purification

```
resMH1 <- difMH(somatic,gender,focal.name=1)
resMH1
```

- Call the MH function with purification

```
resMH2 <- difMH(somatic,gender,focal.name=1,
  purify=TRUE)
resMH2
```

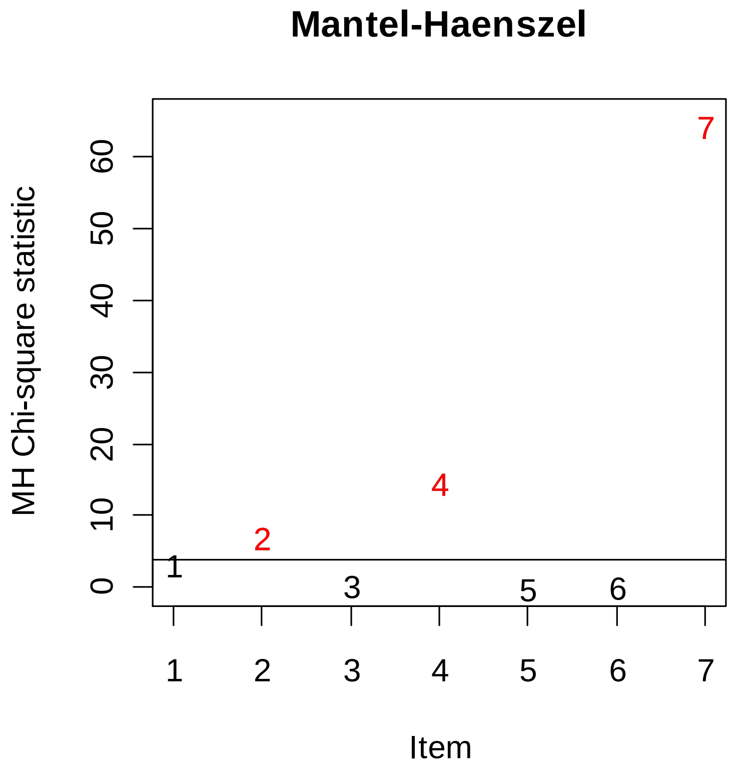
Results for MH chi-square GHQ28 somatic symptoms

```
Detection of Differential Item Functioning using Mantel-Haenszel
   method with continuity correction

Mantel-Haenszel Chi-square statistic:
 without item purification               with item purification
   Stat.   P-value                       Stat.   P-value
V1  3.5263  0.0604 .                     V1  0.4336  0.5102
V2  7.2384  0.0071 **                    V2  1.0218  0.3121
V3  0.5242  0.4690                       V3  0.5919  0.4417
V4 14.8900  0.0001 ***                   V4  8.5337  0.0035 **
V5  0.0818  0.7748                       V5  2.0031  0.1570
V6  0.2905  0.5899                       V6  0.0001  0.9919
V7 64.4739  0.0000 ***                   V7 61.6110  0.0000 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Detection threshold: 3.8415 (significance level: 0.05)
```

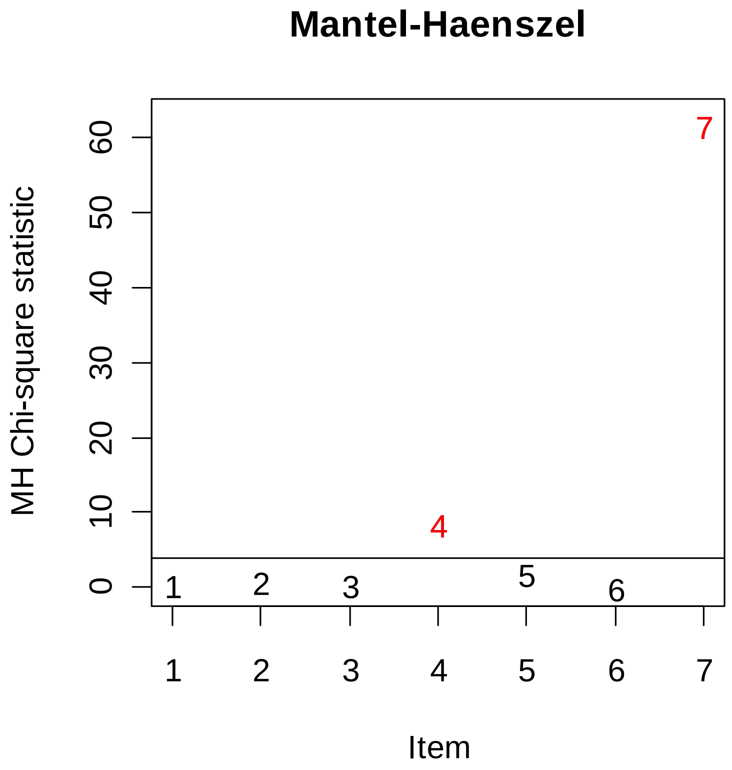
Plotting MH results

- Plots MH chi-square labelled by the item numbers

plot(resMH1)



plot(resMH2)



Results for MH LOR
GHQ28 somatic symptoms

- Log odds ratio is ordered by specifying MHstat="logOR"
  - In fact, its standardized version (z LOR from DIFAS) is printed

without item purification

```
     Stat.   P-value
V1  1.9654  0.0494 *
V2  2.6867  0.0072 **
V3  0.8067  0.4198
V4  3.8930  0.0001 ***
V5 -0.4016  0.6880
V6  0.6571  0.5111
V7 -7.6836  0.0000 ***
```

with item purification

```
     Stat.   P-value
V1  0.7562  0.4495
V2  1.0993  0.2716
V3 -0.8759  0.3811
V4  2.9507  0.0032 **
V5 -1.5249  0.1273
V6  0.1214  0.9034
V7 -7.4617  0.0000 ***
```

Compare Mantel-Haenszel chi-square from R and DIFAS

difR

	Stat.	P-value
V1	3.5263	0.0604 .
V2	7.2384	0.0071 **
V3	0.5242	0.4690
V4	14.8900	0.0001 ***
V5	0.0818	0.7748
V6	0.2905	0.5899
V7	64.4739	0.0000 ***

DIFAS

Name	MH CHI
Var 1	3.5263
Var 2	7.2384
Var 3	0.5242
Var 4	14.8900
Var 5	0.0818
Var 6	0.2905
Var 7	64.4749

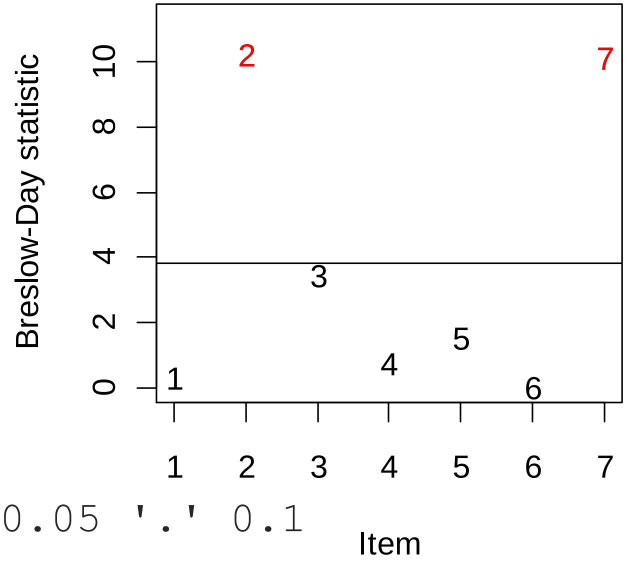
Breslow-Day statistics in **difR**: `difBD`

- Breslow-Day statistic function and its options in difR:

```
resBD <-difBD(somatic,gender,focal.name=0,purify=FALSE,
   BDstat="trend")
```

  - "trend" option uses the same statistic as DIFAS (default is "BD")

```
Breslow-Day trend statistic without item purification:
    Stat.   P-value
V1   0.3691  0.5435
V2  10.3261  0.0013 **
V3   3.5318  0.0602 .
V4   0.8470  0.3574
V5   1.6367  0.2008
V6   0.0799  0.7774
V7  10.2077  0.0014 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

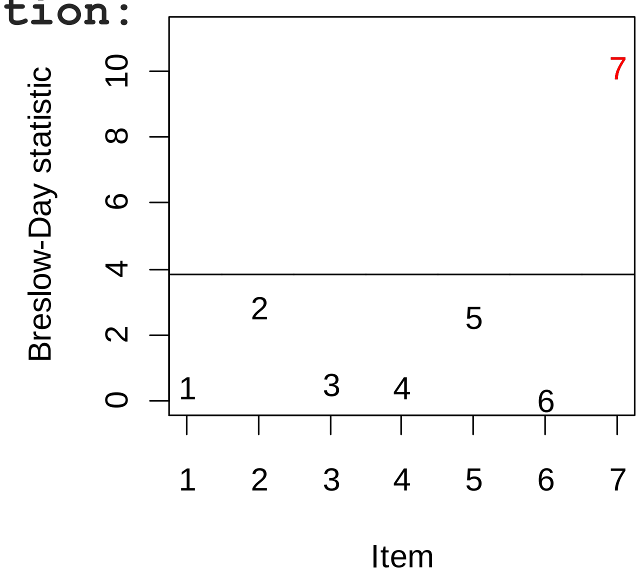


Breslow-Day statistics in **difR**

- Breslow-Day statistic function with purification

```
resBD <- difBD(somatic,gender,focal.name=0,purify=TRUE,
   BDstat="trend")
```

```
Breslow-Day trend statistic with purification:
      Stat.     df       P-value
   V1  0.4777  0.4894
   V2  2.9069  0.0882 .
   V3  0.5615  0.4537
   V4  0.4590  0.4981
   V5  2.6292  0.1049
   V6  0.0774  0.7808
   V7 10.2077  0.0014 **

   Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```



Practical 2.
INSTRUCTIONS for MH DIF in difR

- Now test ‘dichotomousDIF.txt’ for DIF with respect to gender using MH function in **difR**

We have found DIF. What now?

HOW TO DEAL WITH DIF

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 Prof. Ronald Hambleton (his lecture on DIF):
  - 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.

Partial invariance - problem

From Millsap’s lecture at conference “Factor Analysis at 100” (2004)

Example: Suppose that we have p=10 observed measures, with 6 of the 10 measures having invariant loadings. What is the implication of partial invariance for use of the scale formed by the 10 measures? The literature provides little guidance here.

(1) “Go ahead and use the full 10-measure scale because the majority of the measures are invariant.”

--this option ignores the magnitudes of the violations of invariance.

(2) “Go ahead and use all measures as long as none of them show loading differences in excess of ____.”

--this option uses arbitrary standards for deciding when a difference is “too large”.

(3) “Drop any measures that aren’t invariant, and use the remaining measures.”

--this option results in as many versions of the scale as there are invariance studies.

(4) “Don’t use the scale!”

--this option leads to paralysis, or early retirement.

Partial invariance - solution

- **Solution:** Consider whether the violations of invariance interfere with the intended use of the scale (Millsap & Kwok, 2004)

Step 1: Arrive at fitted model for all groups with partial invariance.

Step 2: Use fitted model to generate hypothetical bivariate distribution of factor scores pooled across groups.

Step 3: Designate cut points on factor score distributions for selection.

Step 4: Calculate sensitivity, specificity, hit rate for each group, and compare to strict invariance model.

Step 5: Base decision on above accuracy indices.

How to adjust for DIF

- It is possible to adjust for DIF in the model
  - Use partially invariant model for scoring
  - For example, can release parameter constraints between the groups in Mplus
- Crane et al. (2004, 2006)

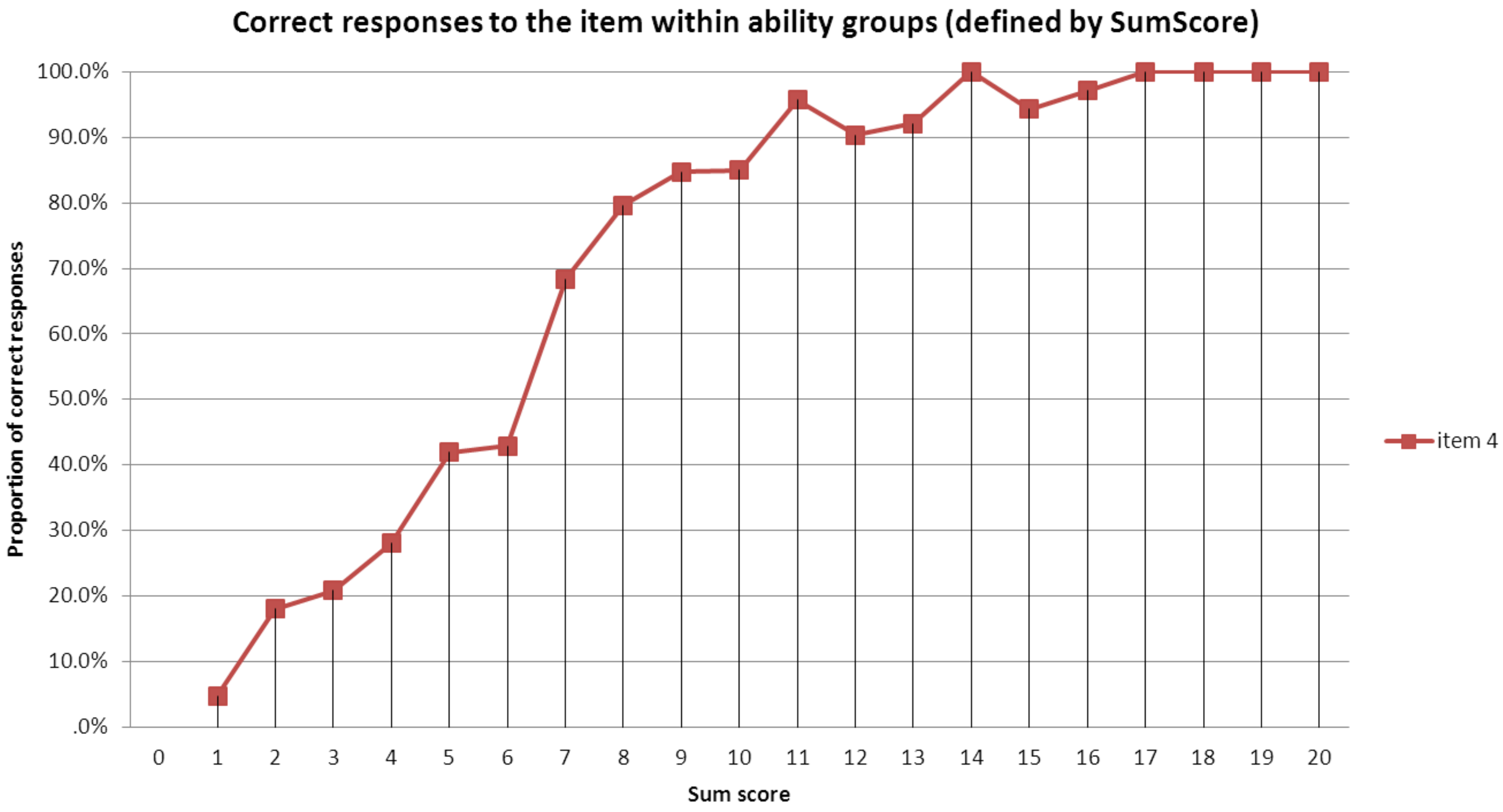
a) items without DIF have item parameters estimated from whole sample – (anchors)

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

Parametric methods for detecting DIF

LOGISTIC REGRESSION METHOD

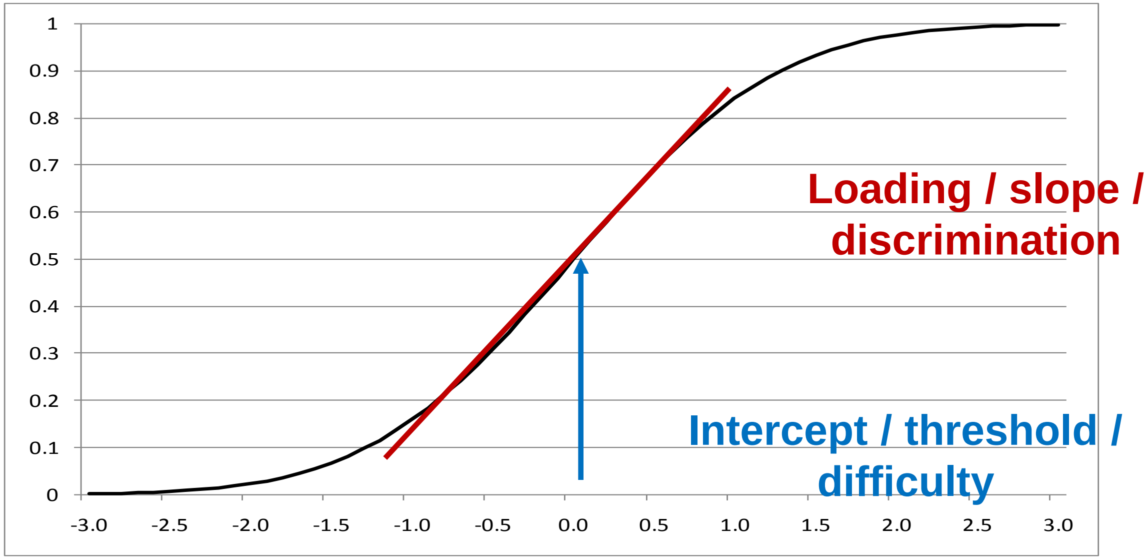
Likelihood of correct response as function of ability



Logistic Regression to detect DIF

- It is assumed that you have a proxy for the latent construct
  - – sum score, estimate of ability from an IRT model…
- Empirical relative frequencies of endorsing an item depending on this proxy should approximately follow an s-shaped curve
  - – in IRT it is called Item Characteristic Curve or ICC

Parametric ICC

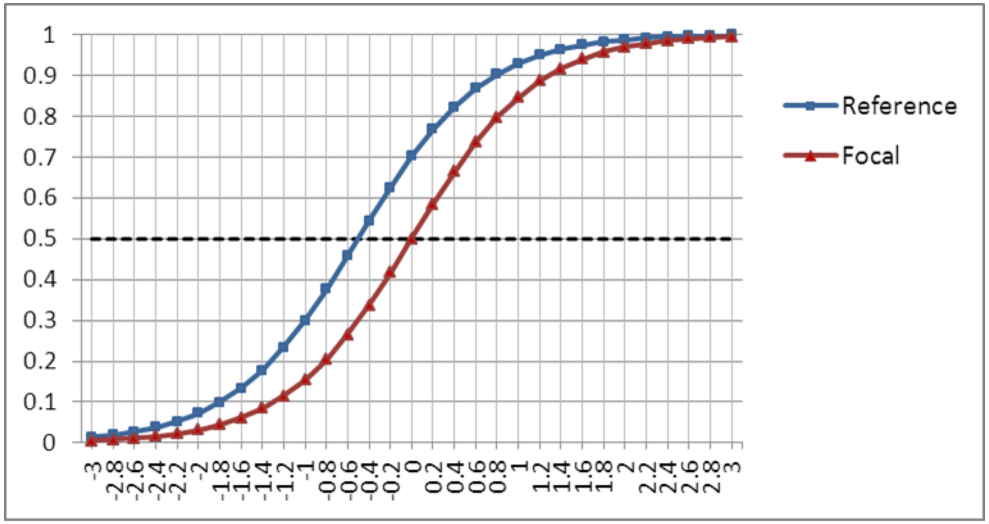


- For binary items, relationship between probability of correct response and the latent attribute is described by logistic regression

$$P(X_i = 1) = \frac{e^{\alpha+\beta F}}{1+e^{\alpha+\beta F}}$$

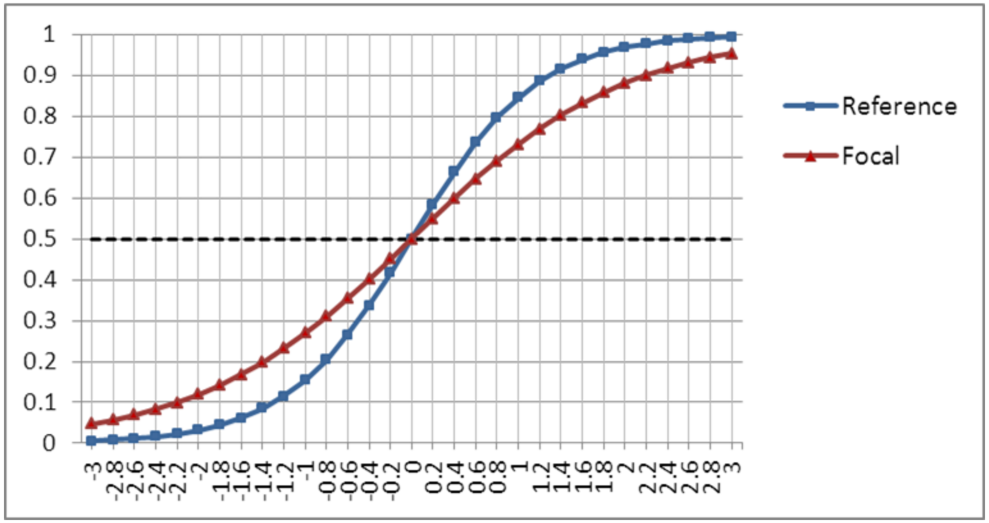
$$\log^e\left(\frac{P(X_i)}{1-P(X_i)}\right) = \alpha + \beta F$$

Uniform and non-uniform DIF



UNIFORM

Focal and Reference groups have different intercept (alpha) parameters



NON-UNIFORM

Focal and Reference groups have different slope (beta) parameters

Logistic Regression to detect DIF

1. Test a baseline model that predicts the probability of a correct answer from the level of the attribute

$$\ln\left(\frac{P(X_i=1)}{1-P(X_i=1)}\right)=\alpha+\beta F$$

2. Add a grouping variable into the regression to see if there is any uniform DIF

$$\ln\left(\frac{P(X_i=1)}{1-P(X_i=1)}\right)=\alpha+\beta F+\mathrm{c}\cdot group$$

3. Add an interaction term between attribute and group to see if there is any non-uniform DIF

$$\ln\left(\frac{P(X_i=1)}{1-P(X_i=1)}\right)=\alpha+\beta F+\mathrm{c}\cdot group+\mathrm{d}\cdot F\cdot group$$

- Quantify the significance and the effect size of each step

Testing Logistic Regression models

1. Improvement in chi-square fit in Model 2 against Model 1 is tested
   - 1 degree of freedom
   - If adding the grouping variable significantly improved the fit, then uniform DIF might be present.
2. Improvement in chi-square fit in Model 3 against Model 2 is tested
   - 1 degree of freedom
   - If adding the interaction term significantly improved the fit, then non-uniform DIF might be present.

Associated effect sizes

- Zumbo-Thomas (1997) – too lenient in most cases
    - The item is displaying DIF if p-value <= 0.01 and R-squared > 0.13
- Gierl & McEwen (1998), Jodoin (1999) – more conservative criteria
    - Large or C-level DIF: R-squared ≥ 0.07 **AND** chi square significant
    - Moderate or B-level DIF: R-squared between 0.035 and 0.07; **AND** chi square significant
    - Negligible or A-level DIF: R-squared < 0.035 **OR** chi square insignificant

Logistic Regression in **difR**

- Logistic regression with *score* as predictor

```
difLogistic(Data,group,focal.name,
criterion="LRT", type="both", alpha=.01,
purify=TRUE, plot="lrStat")
```

  - Type can be "**both**" (default), "**udif**", "**nudif**"
  - Criterion can be "**LRT**" (likelihood ratio test, default) or "**Wald**"
  - Plot can be "**lrStat**" (default) or "**itemCurve**"

Results for the GHQ-28 Somatic symptoms (uniform DIF)

```
resLR1 <- difLogistic(somatic, gender, focal.name=1,
   purify=TRUE, type="udif")
```

```
LR DIF statistic:
        Stat.   P-value
V1     0.2079   0.6484
V2     1.5597   0.2117
V3     0.3168   0.5736
V4    11.8350   0.0006 ***
V5     0.8005   0.3709
V6     0.7212   0.3958
V7    81.1281   0.0000 ***
```

```
Effect size (Nagelkerke's R^2):
   R^2      ZT JG
V1 0.0000 A  A
V2 0.0000 A  A
V3 0.0000 A  A
V4 0.0081 A  A
V5 0.0012 A  A
V6 0.0010 A  A
V7 0.0732 A  C
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Detection threshold: 3.8415 (significance level: 0.05)
```

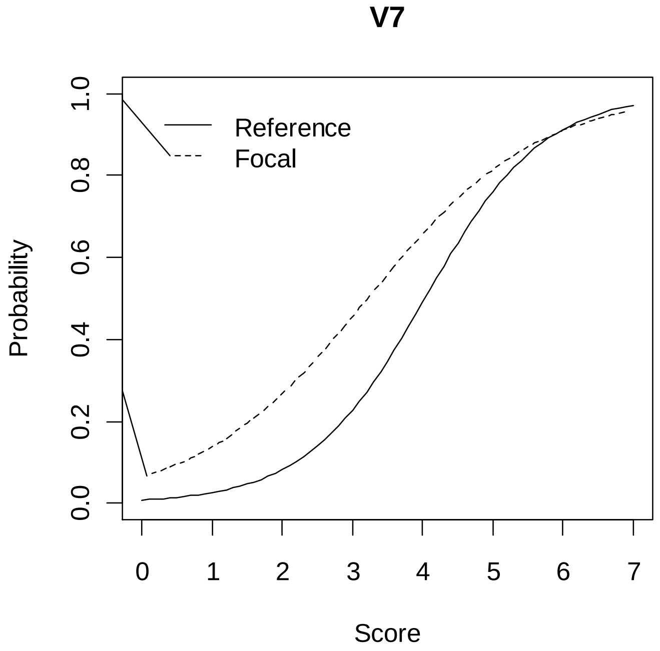
Results for the GHQ-28 Somatic symptoms (non-uniform DIF)

```
resLR2 <- difLogistic(somatic, gender, focal.name=1,
   purify=TRUE, type="nudif")
```

```
LR DIF statistic:                Effect size (Nagelkerke's R^2):
     Stat.    P-value             R^2      ZT JG
V1   1.1483   0.2839              V1 0.0000 A   A
V2   0.1086   0.7417              V2 0.0000 A   A
V3   0.1687   0.6813              V3 0.0000 A   A
V4   0.1122   0.7376              V4 0.0001 A   A
V5   1.1704   0.2793              V5 0.0019 A   A
V6   0.0275   0.8684              V6 0.0000 A   A
V7  12.7631   0.0004 ***          V7 0.0120 A   A

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Detection threshold: 3.8415 (significance level: 0.05)
```

Plot ICCs for item 7

- `resLR3 <- difLogistic(somatic, gender, focal.name=1, purify=TRUE, type="both")`



- `plot(resLR3,plot= "itemCurve", item=7)`

Probing multiple criteria

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

```
resDIF<-dichoDif(somatic, gender, focal.name=1,
  method=c("MH", "Logistic", "BD"), purify=TRUE)
```

```
Comparison of DIF detection results:

   M-H   Logistic BD    #DIF
V1 NoDIF NoDIF    NoDIF  0/3
V2 NoDIF NoDIF    NoDIF  0/3
V3 NoDIF NoDIF    NoDIF  0/3
V4  DIF   DIF      DIF   3/3
V5 NoDIF NoDIF    NoDIF  0/3
V6 NoDIF NoDIF    NoDIF  0/3
V7  DIF   DIF      DIF   3/3
```

Practical 3.
INSTRUCTIONS for LR DIF in difR

- Use difR's logistic regression function to test our ability items (‘dichotomousDIF.txt’) for DIF with respect to gender

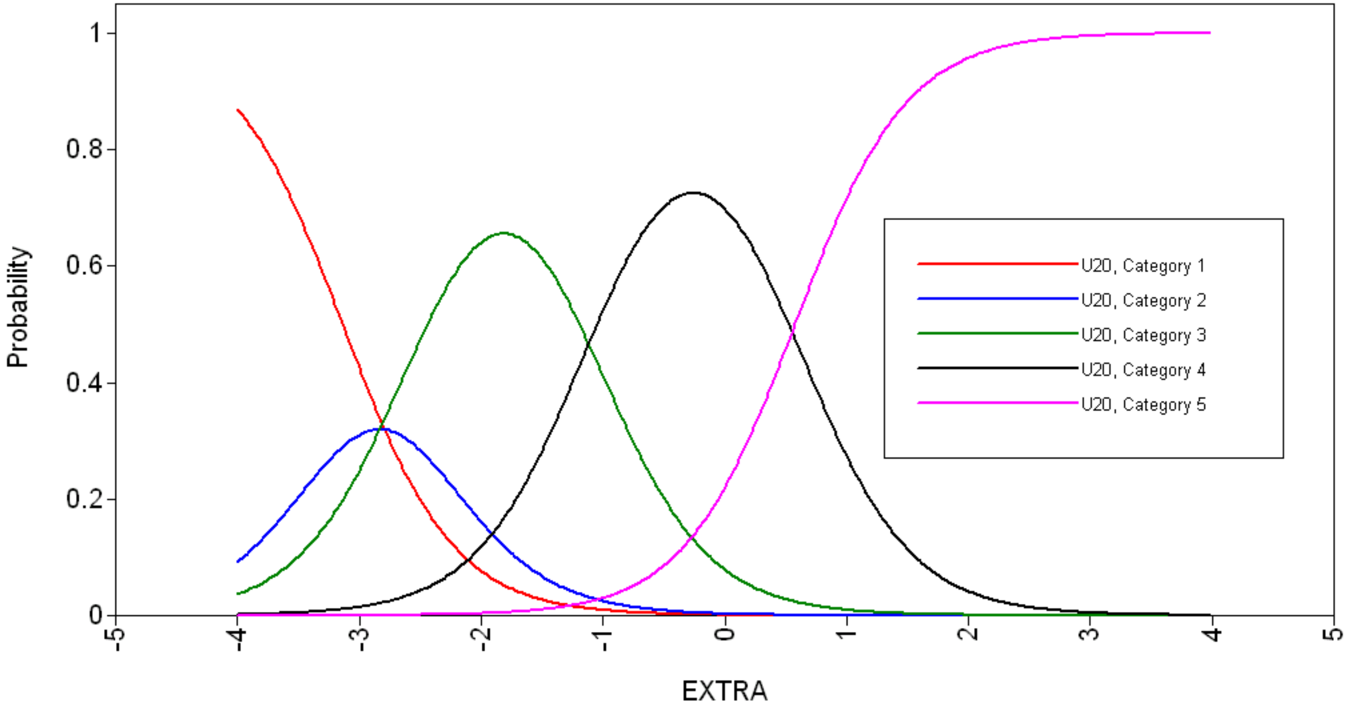
1. **X** fits an item response model

ORDINAL ITEMS

ICCs for an ordinal item

- An example item with 5 response categories
  - such as

  'strongly disagree' - 'disagree' - 'neutral' - 'agree' - 'strongly agree'



Extending the MH statistic to ordinal items

- Mantel’s (1963) chi-square test (not an extension of the MH test) can be used with polytomous items
  - distributed as chi-square with one degree of freedom.
- Liu and Agresti (1996) extended the MH statistic for use with ordinal variables
  - It is a generalization of the MH common odds ratio
  - Is asymptotically normally distributed.
- Penfield and Algina (2003) applied the Liu Agresti estimator to detect DIF in polytomous items
  - They provide computational detail
  - it is interpreted in the same frame of reference as the MH common odds ratio
  - fully implemented in DIFAS

Examining gender DIF for GHQ28 somatic symptoms

Lui-Agresti common log-odds ratio

```
DIF STATISTICS: POLYTOMOUS ITEMS
---------------------------------------------------------------------------
Name          Mantel      L-A LOR    LOR SE     LOR Z     COX'S B  COX SE    COX Z
---------------------------------------------------------------------------
Var 1         31.499      0.592      0.107      5.533     0.561    0.1       5.61
Var 2         29.325      0.488      0.09       5.422     0.429    0.0792    5.417
Var 3         29.144      0.52       0.097      5.361     0.476    0.0882    5.397
Var 4         94.930      1.026      0.105      9.771     0.828    0.0849    9.753
Var 5          2.031     -0.142      0.101     -1.406    -0.122    0.086    -1.419
Var 6          2.157      0.162      0.113      1.434     0.135    0.0917    1.472
Var 7        352.777     -1.827      0.106    -17.236    -1.177    0.0626  -18.802
---------------------------------------------------------------------------
Reference Value = 0, Focal Value = 1
```

LOR Z is greater than 1.96 in magnitude, indicating highly significant DIF

Differential Step Functioning (DSF)

- Examination of DSF effects can prove useful in understanding the location of the DIF effect (i.e., which response option(s) manifest the DIF effect).
- Dichotomisation is performed for these analyses; each trace line is considered through either
  - The cumulative approach (cumulative DIF effect as category increases)
  - The adjacent categories approach (category-specific DIF effect)

Examining gender DSF for GHQ28 somatic symptoms

Cumulative

```
DSF for Var 7
------------------------------------
Step      CU-LOR   SE          Z
------------------------------------
2         -2.05    0.10543    -19.444
3         -1.378   0.17395    -7.922
4         -1.357   0.42955    -3.159
------------------------------------
L-A LOR = -1.827   LOR SE = 0.106
```

Adjacent categories

```
DSF for Var 7
------------------------------------
Step      AC-LOR    SE         Z
------------------------------------
2         -1.947   0.112      -17.384
3         -0.473   0.20314    -2.328
4         -0.519   0.48286    -1.075
------------------------------------
Item-Level LOR = -1.555
```

It appears that the second category (“no more than usual”) has the greatest effect

Extending logistic regression to ordinal items

- **lordif** package in R can perform LR for polytomous items

```
library(lordif)
```

- Procedure `lordif()`
  - Matching variable is the latent trait score estimated by the IRT method
  - Purification is always performed

```
lordif(resp.data, group,
criterion = c("Chisqr", "R2", "Beta"),
pseudo.R2 = c("McFadden", "Nagelkerke",
"CoxSnell"), alpha = 0.01)
```

Examining gender DIF for GHQ28 somatic symptoms with LR

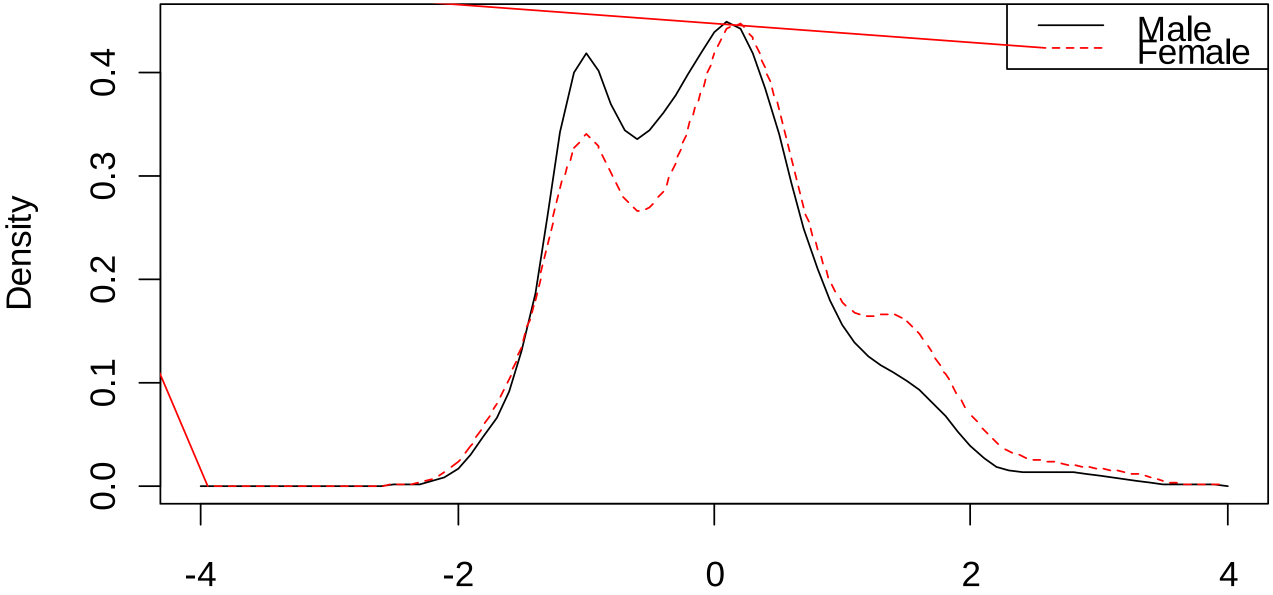
```
Number of items flagged for DIF: 1 of 7
  Items flagged: 7

  Threshold: R-square change = 0.02
Item    ncat       12          13          23
--------------------------------------------
1       4      0.0009      0.0032      0.0023
2       4      0.0000      0.0001      0.0001
3       4      0.0002      0.0002      0.0000
4       4      0.0038      0.0043      0.0004
5       3      0.0068      0.0069      0.0001
6       3      0.0019      0.0022      0.0003
7       4      0.0990      0.1061      0.0071
```

- `1-2`: Model 2 compared to baseline model (test for uniform DIF)
- `1-3`: Model 3 compared to baseline model (test for general DIF)
- `2-3`: Model 3 compared to Model 2 (test for non-uniform DIF)

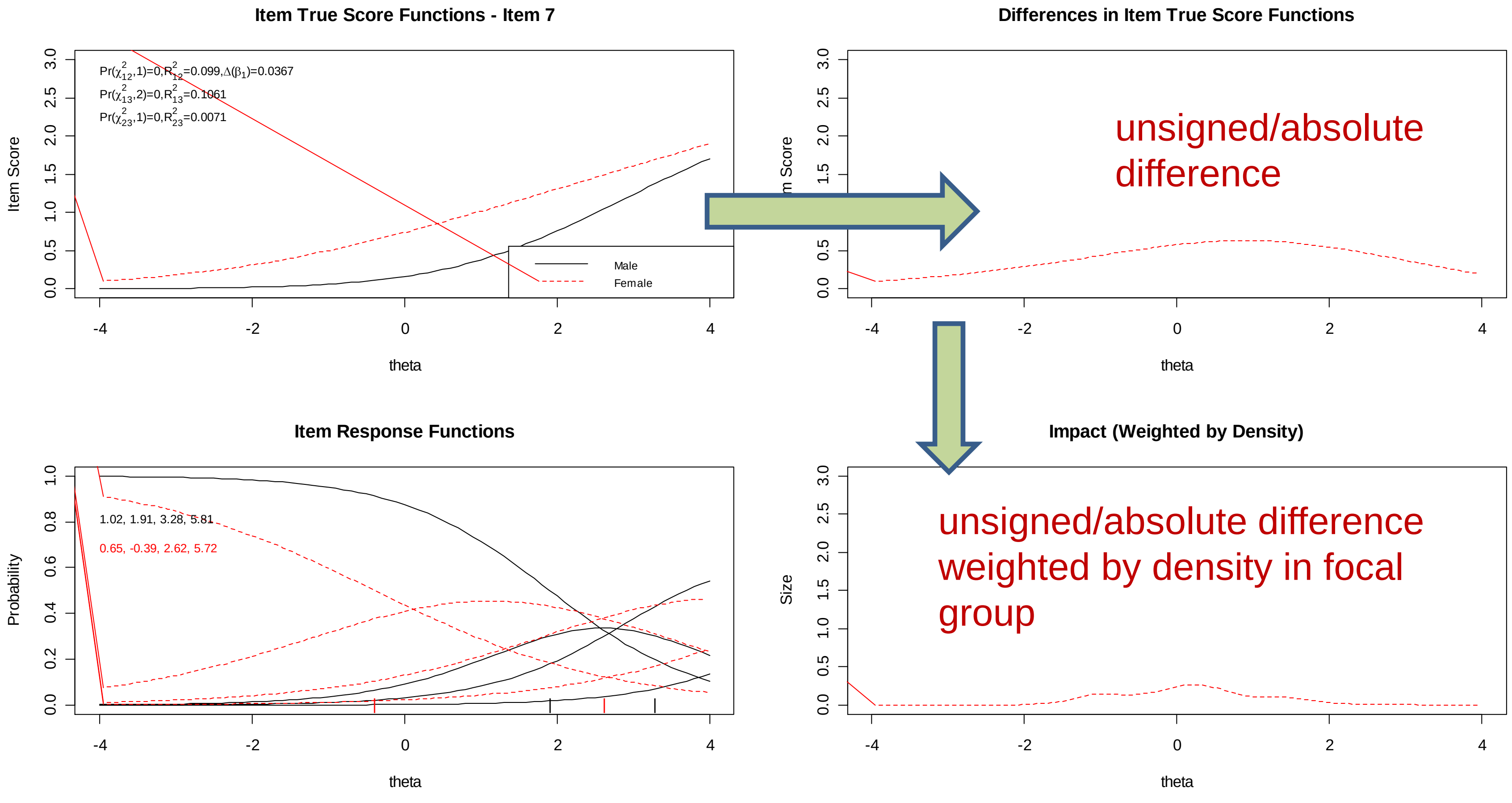
Plot the trait distributions

```
plot(resLR, labels = c("Male","Female"))
```

- Latent construct distributions of reference and focal groups
    - In this case, distributions are quite similar



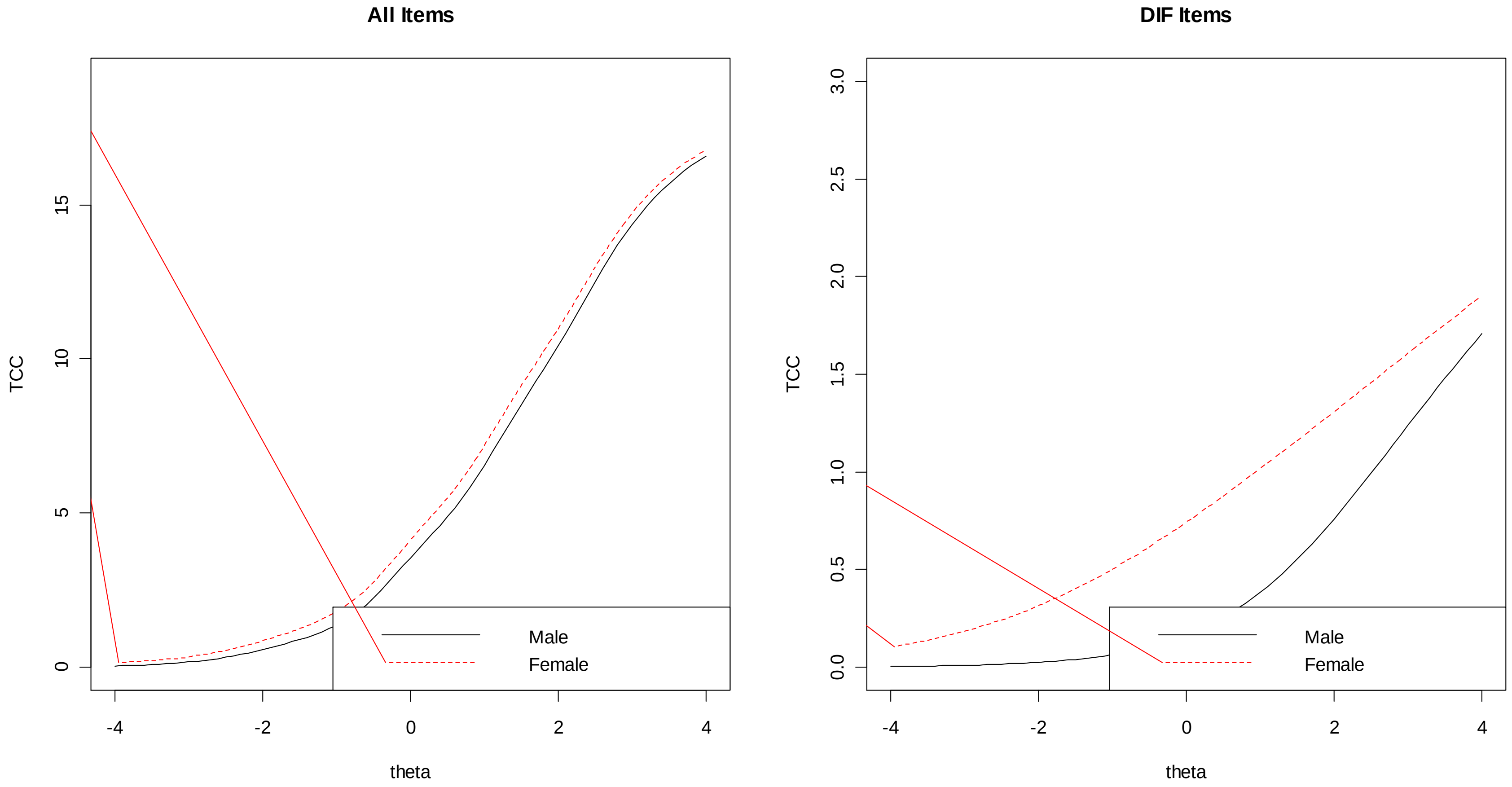
Plot the DIF item trace lines



unsigned/absolute difference

unsigned/absolute difference weighted by density in focal group

Plot the test characteristics curves



2. **X** fits a common factor model

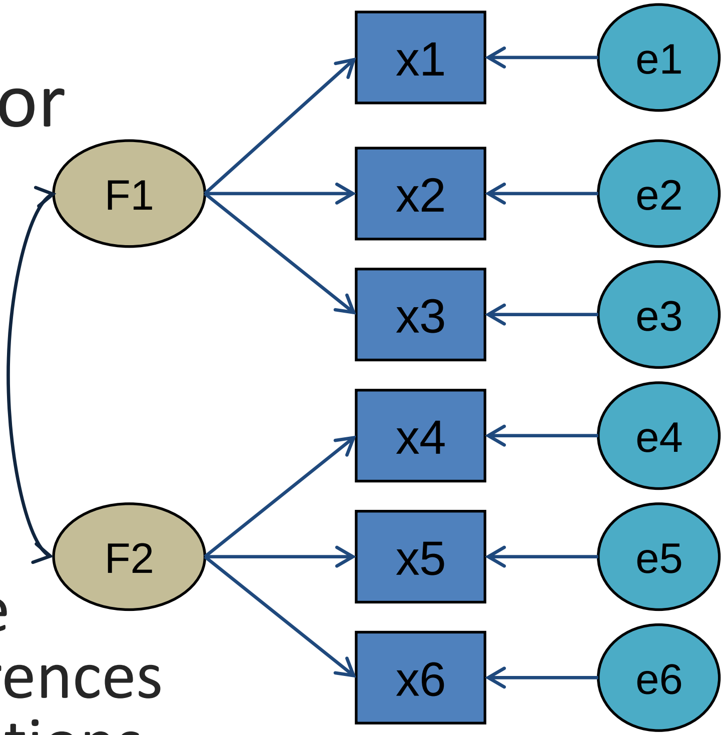
CONTINUOUS VARIABLES

Factorial invariance

- Factor model (2-dimensional for example)

  $$\mathbf{X}_k = \boldsymbol{\mu}_k + \Lambda_k F1_k + \Lambda_k F2_k + \mathbf{e}_k$$

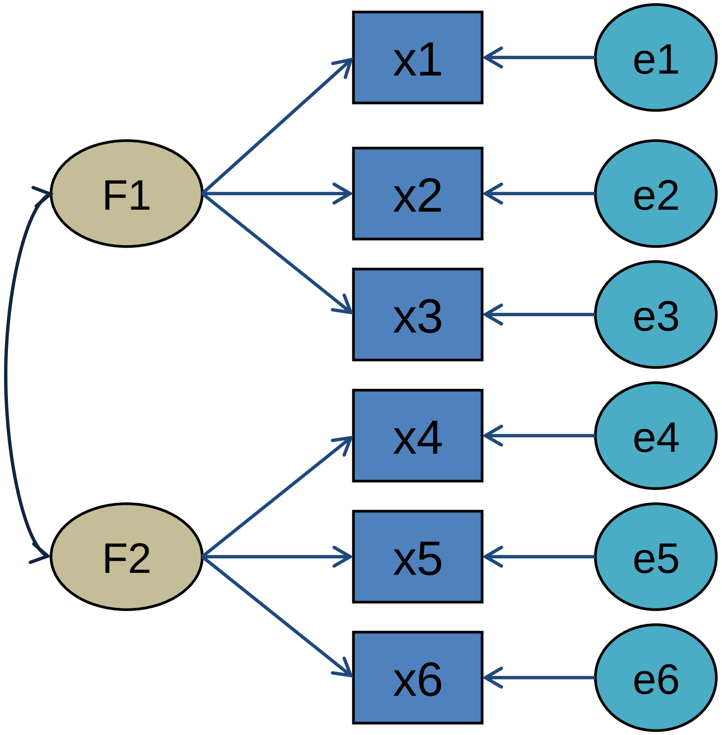
  – *k* is a group indicator



- Factorial invariance
  - Systematic group differences in observed means and covariance matrices are due to group differences in common factor score distributions.
  - Invariance in the factor means or covariances is not required.

Elements of the common factor model

1. the model specification (number of factors and loading pattern),
2. the regression coefficients,
3. the regression intercepts,
4. the regression residual variances,
5. the means of the common factors,
6. the variances of the common factors, and
7. the covariances among the common factors.



- The last 3 elements are not considered necessary for MI to hold

Studying factorial invariance

Typically, invariance is studied via a nested sequence of models:

(1) **Configural invariance** (Thurstone, 1947): zero elements of pattern matrices in the same locations for all groups.

(2) **Metric or pattern invariance** (Thurstone, 1947): pattern matrices are fully invariant.

(3) **Strong factorial invariance** (Meredith, 1993): pattern matrices and latent intercepts are fully invariant.

(4) **Strict factorial invariance** (Meredith, 1993): pattern matrices, intercepts, and unique variances are fully invariant.

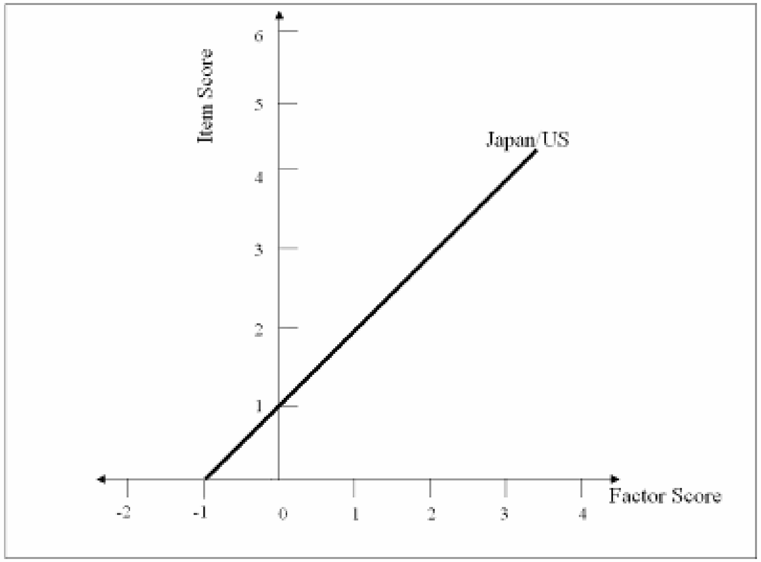
Meredith argued that strict invariance is a necessary condition for a fair and equitable comparison. Unfortunately, it rarely holds.

Strict factorial invariance

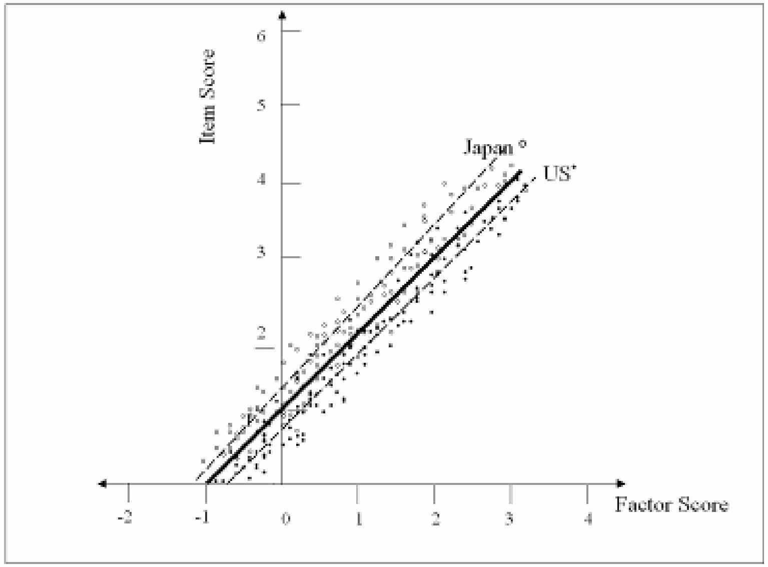
- Vandenberg and Lance (2000) review of published MI studies
  - 99% of the MI studies investigated loading invariance
  - 12% investigated intercept equality
  - 49% investigated residual variance equality.
- Equality in all 4 elements is necessary for MI
- Strict factorial invariance would ensure that the relationship between the factors and the observed item scores remain the same across groups

Violations of strict invariance -1

Strict factorial invariance



Unequal item-specific effects

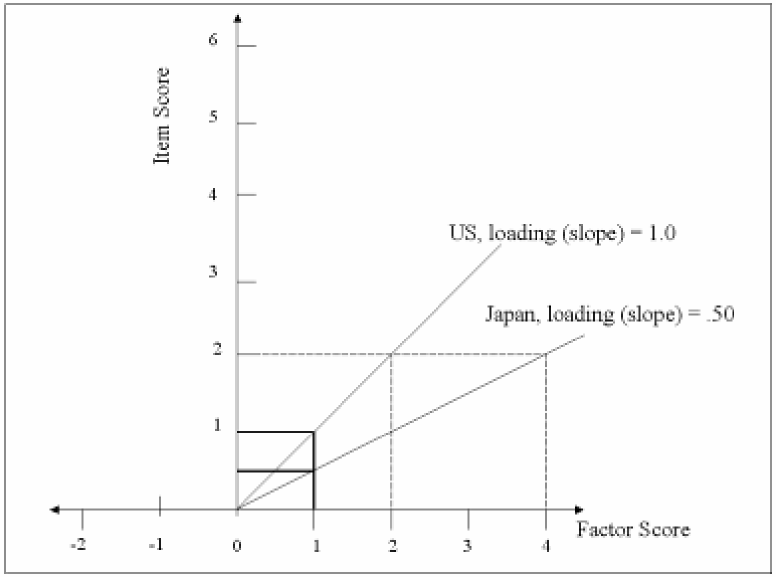


Residuals are systematically higher for Japan (indicated by “○”) then those of the U.S. (indicated by “●”), as is the variation among the Japanese respondents

Illustration from Wu, Li, and Zumbo (2007)

Violations of strict invariance - 2

Unequal factor loadings



Unequal intercepts

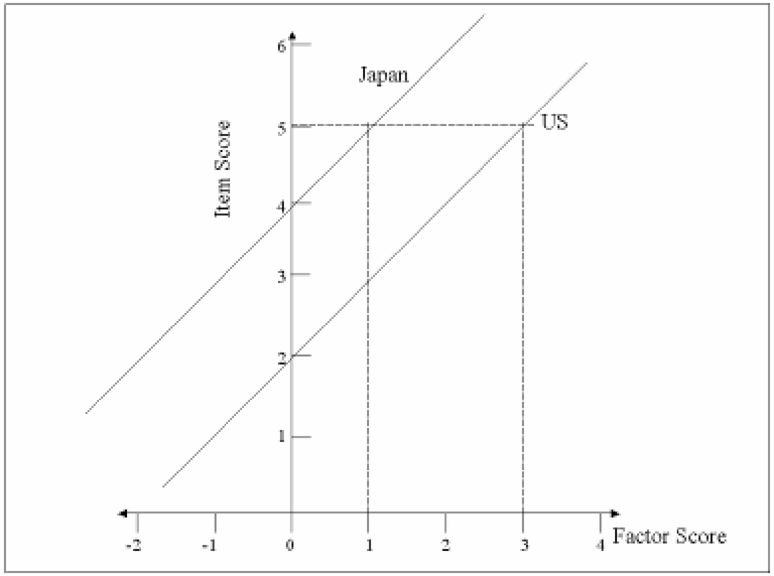


Illustration from Wu, Li, and Zumbo (2007)

Levels of measurement invariance

- Van de Vijver & Poortinga (1997) define levels of MI
  - Structural / functional invariance
    - The same psychological constructs are measured across groups
    - This is ensured by **Metric or pattern invariance**
  - Measurement unit invariance
    - The same measurement unit (individual differences found in group A can be compared with differences found in group B)
    - Factor loadings and residual variances should be the same (intercepts can be different!)
  - Scalar / full score invariance
    - The same measurement unit and the same origin (scores can be compared across groups)
    - Factor loadings, intercepts and residual variances should be the same. This is ensured by **Strict factorial invariance**

Influence of bias on the level of invariance

Type of Bias	Structural	Measurement unit	Scalar
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, 1997

Multi-group CFA

- MG-CFA is the most widely used method for investigating factorial invariance
- Factor models are specified in two groups and a series of equality constraints are tested
- Alternative strategies:
  - Start with free model and add constraints (e.g. from configural invariance to strict invariance). Stop when model does not fit the data any longer.
  - Start with fully constrained model and release constraints until the model fits the data.

Identification in MG-CFA

- Identification in the first group is the same as in one-group analysis
  - the model in the second group can be identified through constraints (e.g. factor mean and variance can be freely estimated)
- Factor models are identified either by fixing one item’s loading, or factor variance (and, one item’s intercept or factor mean)
- What if the item chosen is biased???
  - Finding "referent" item (item with no bias)

Strategy 1 (adding constraints)

(1) Start with the same configural model *using a **referent item** for identification*, and no constraints on other parameters (configural invariance)

(2) Sequentially add the invariance constraints on factor loadings (pattern invariance).

(3) Continue adding the constraints until either fit is inadequate or all loadings are constrained.

(4) Repeat these steps with intercepts (strong invariance), confining interest to measures that have invariant loadings.

(5) Repeat these steps with residual variances (strict invariance), confining interest to measures that have invariant loadings and intercepts.

The problem with this strategy is that it is very labour-intensive. Also, how to find referent (DIF-free) item?

Finding the referent item

From Stark, Chernyshenko & Drasgow (2006)

1. All slopes and intercepts are constrained to be equal across groups. The mean of the reference group is set to 0, the variance is set to 1; and the second group's mean and variance are free.
2. Run a series of augmented (partially constrained) models for each item. In an augmented model for item 1 , everything is constrained equal across groups apart from slope and intercept for item 1, which are free. Do that for all items and record chi-square changes in relation to the fully constrained model.
3. An item "wins" this race if 1) it has insignificant change chi-square; 2) it has the highest slope out of all items. This is the **referent item**.

- Slope and intercept are tested together (uniform and non-uniform DIF), because the reference item has to have neither.

Strategy 2 (relaxing constraints guided by modification indices)

The criticism of this strategy is that the comparison of nested models is not proper if there are violations of MI; particularly if the number of DIF items is large.

- Hernandez, Stark & Chernyshenko (2008) show through simulation studies that the following approach is effective and its power and error rate comparable to Strategy 1

1. Start with fully constrained model where mean and variance of the factor in the first group are set.
2. If the largest MI is statistically significant then fit a new model relaxing the group constraint on that parameter.
3. Evaluate statistical significance of the largest MI associated with the constrained parameters, and modify the model again. This iterative procedure continues until the largest MI is not statistically significant.
4. An item is flagged as showing DIF if there were significant differences in the loading, the intercept, or both parameters.

How to judge fit?

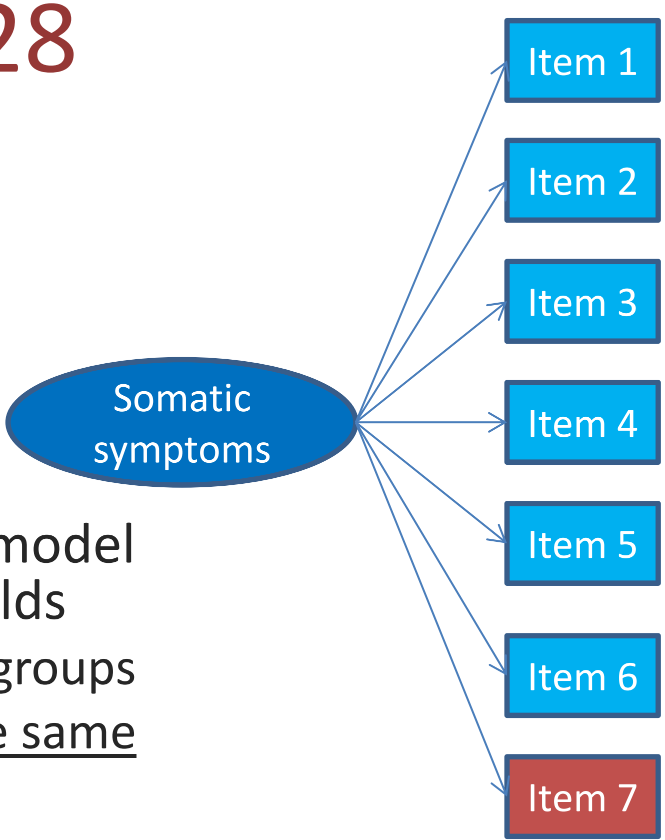
- χ^2 as well $\Delta\ \chi^2$ are affected by sample size and model complexity
- Most relative fit indices have been found to be affected by the model complexity too
- Cheung and Rensvold (2002) conducted a comprehensive study of fit indices
  - only RMSEA was not affected by model complexity and RMSEA≤ 0.05 is recommended for indicating the configural model fit
  - ΔCFI ≤ -0.01, ΔGamma Hat ≤ -0.001, and ΔMcdonald's Non-Centrality Index ≤ -.02 were the best indication of support of MI.

Note on MG-CFA approach for binary and ordinal items

- MG-CFA approach is also applicable to binary and ordinal items (those we model with IRT)
  - Because IRT models are CFA models with categorical variables
  - MG-CFA is a parametric DIF method
    - Parameters are item intercepts and factor loadings
  - For categorical items, no residual variance may be identified.
    - Therefore, all residuals are fixed to 1 in one group
    - They can be freely estimated in the other group; however, it is usual in IRT applications to consider them all equal 1, so that item slope absorbs any inequality

Illustration with NSHD dataset and GHQ-28

- Back to our GHQ-28 Somatic symptoms subscale, and gender-related DIF.
- For purposes of this illustration, we first assume the 4-category item responses continuous.
- We will use M*plus*
- By default, M*plus* sets up a multi-group model so that the strong factorial invariance holds
  - Factor means and variances vary across groups
  - Loadings and intercepts are assumed the same
  - Residuals vary across groups
  - If we want test for strict invariance, we need to constrain the residuals manually



Mplus: a short reminder

- Intercepts and means are referred to as

  `[i1]; [i2]; … [i7];`

  or in expansion format `[i1-i7];`
- Loadings are referred to as

  `Somatic BY i1-i7*;`
- Variances and residual variances are referred to as

  `i1; i2; … i7;`

  or in expansion format `i1-i7;`
- Declare parameter numbers

  `i1-i7  (1-7);`
- Parameter numbers can be used for equality constraints

Mplus syntax for Strict invariance

```
DATA: FILE IS LikertGHQ28_sex.dat;
VARIABLE: NAMES ARE i1-i28 gender;
   USEVARIABLES ARE i1-i7;
   GROUPING IS gender (1=female, 0=male);
ANALYSIS: !defaults are fine, this section is empty
MODEL:
   Somatic BY i1-i7*;
MODEL male:
   i1-i7 (1-7);
   Somatic@1; [Somatic@0];
MODEL female:
   i1-i7 (1-7);
OUTPUT:  MODINDICES(10);
```

Factor is indicated BY items 1 to 7; all loadings are freely estimated

Set parameter numbers for residuals in male group

Refer to the same parameter numbers in female group – this will ensure that the residuals are equal

Examining modification indices

```
                    M.I.             E.P.C
----------------------------------------------
WITH Statements
I6   WITH I5       347.803       0.128


Variances/Residual Variances
I7                 206.258      -0.184



Means/Intercepts/Thresholds
[ I7]              395.942      -0.252
```

Item 5 and item 6 share common variance after controlling for somatic symptoms

Note large MI for intercept and residual of item 7

Fit indices for nested models

Condition	Chi-square	CFI	RMSEA
Strict invariance	2498 (df=47)	.669	.190
Strict invariance, with item-parcel for items 5 and 6	1258 (df=34)	.791	.158
Intercept for i7 released (uniform DIF)	832 (df=33)	.864	.129
Residual for i7 released (item-specific variance)	513 (df=32)	.918	.102
Residual for parcel 56 released	466 (df=31)	.926	.098

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

Examining final outputs: invariant parameters

```
                    Estimate      S.E.
SOMATIC  BY
I1                   0.308       0.010
I2                   0.558       0.015
I3                   0.615       0.015
I4                   0.461       0.013
PARCEL56             0.382       0.018
I7                   0.211       0.012

Intercepts
I1                   2.035       0.012
I2                   1.739       0.018
I3                   1.715       0.018
I4                   1.414       0.016
PARCEL56             2.524       0.020
```

Examining final outputs: factor means and variances, and the DIF item

males

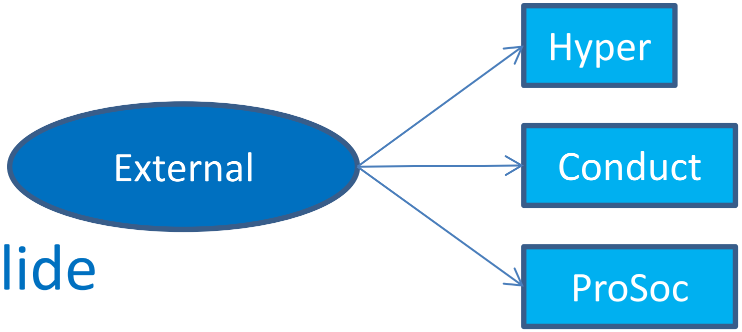
```
                 Estimate  S.E.
Means
 SOMATIC          0.000     0.000
Variances
 SOMATIC          1.000     0.000

Intercepts
 I7               1.200     0.014
Residual Variances
 I7               0.222     0.009
 parcel56         0.675     0.026
```

females

```
                 Estimate   S.E.
Means
 SOMATIC          0.229      0.044
Variances
 SOMATIC          1.402      0.085

Intercepts
 I7               1.718      0.021
Residual Variances
 I7               0.591      0.022
 parcel56         0.983      0.037
```

Practical 4. SDQ Externalising

- Test gender invariance with the multi-group approach in Mplus for the “Externalising” construct based on SDQ
- Strength and Difficulties Questionnaire (SDQ; Goodman); designed to screen children with mental health problems
- 3 subscales form an “Externalising problems” factor
  - Hyperactivity (+), Conduct problems (+), Pro-social behaviour (-)
- Pupils of year 7 (11 years old):
  - 2545 boys and 2794 girls
- Data is in “SDQpupil.dat”
- Variables are described in the next slide



Practical 4. Variables description for SDQ dataset

```
DATA: FILE IS SDQpupil.dat;

VARIABLE: NAMES ARE
hyper emot cond peer pros impact total
hyper2 emot2 cond2 peer2 pros2 impact2 total2
gender;

USEVARIABLES ARE hyper cond pros;
MISSING ARE ALL .;
GROUPING IS gender (1=female, 0=male);
```

Mplus syntax for the baseline model; SDQ Externalising

```
MODEL:
external BY hype* cond pros;

MODEL male:
   [external@0];
   external@1;
   hype-pros (1-3);

MODEL female:
   hype-pros (1-3);

OUTPUT:   MODINDICES(10);
```

SDQ modelling results

Condition	Chi-square	CFI	RMSEA
Strict invariance	312 (df=7)	.901	.128
Intercept for Pro-social scale released	139 (df=6)	.957	.091
Residual for Pro-social scale released	61 (df=5)	.982	.065

SDQ Externalising : final outputs

boys

```
Estimate       S.E.
Means
 External    0.000      0.000
Variances
 External    1.000      0.000

Intercepts
 PROS        6.992      0.039
 Residual Variances
 PROS        3.114      0.094
```

girls

```
Estimate       S.E.
Means
 External  -0.516      0.030
Variances
 External   0.684      0.034

Intercepts
 PROS       7.648      0.039
Residual Variances
 PROS       2.163      0.064
```

Special case: MI with repeated measures

LONGITUDINAL INVARIANCE

Measurement invariance assumptions

- In longitudinal measurement we implicitly make an assumption that our tests measured the same construct(s) across the time points
  - So we assume that the factor loadings, thresholds and residuals stay the same
- Is this a fair assumption to make?
- Does this assumption hold?

Example: measuring self-esteem

- Measuring self-esteem longitudinally (from Horn, 1991)
- Suppose our measure is:
  - Do you feel you are as good looking as the average person?
  - Do you feel you are every bit as smart as the average person?
  - Do you feel you are liked by others as much as the average person is liked?
- Would the concept “self-esteem” have the same meaning (construct validity) for 20-year olds and for 60-year olds?

Example – continued

- Factor patterns might be like this

  Self-esteem = .6*looks + .3*smart + .4*likable

  Self-esteem = .0*looks + .8*smart + .4*likable

  – Guess which one might be found in a sample of 20-year olds?
- Qualitative difference in what is being measured
- We cannot simply sum these items to produce a valid measure of self-esteem in a longitudinal design

Why is DIF important in longitudinal measurement?

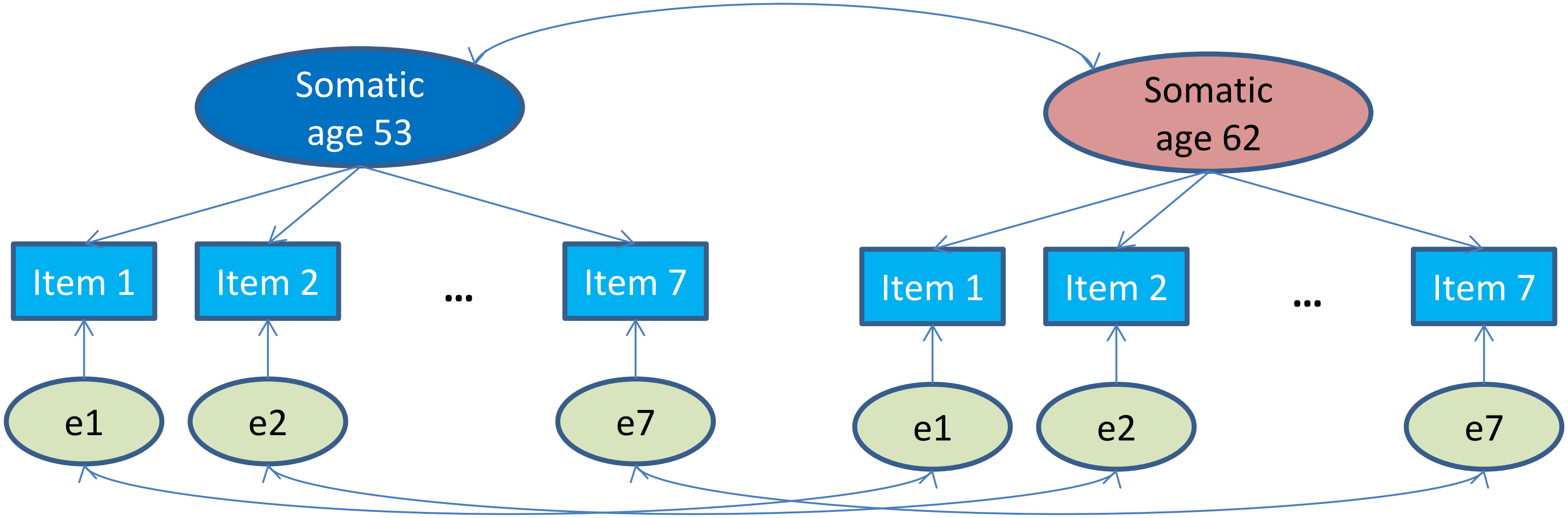
- For example, difference score between T2 and T1 (treatment effect etc.) relies on measurement equivalence at the item level
  - Equal thresholds (no uniform DIF)
  - Equal loadings (no non-uniform DIF)
- If goes unnoticed, DIF distorts the model results
  - We can mistakenly take uniform DIF for real change in the construct level
  - Or non-uniform DIF for reduced stability of the construct

Setting up a longitudinal invariance model

- This is no longer a multi-group model
  - One group, repeated measures
- MI holds if all loadings, intercepts and residuals are the same across time
  - Factor mean and variance can vary across time
  - Factor mean is set to 0 and variance to 1 at T1, and freely estimated at T2
- Important feature of repeated measures is that item residuals might not be independent across time
  - Specific variance in items may be sustained over time
  - For instance, tendency for headaches regardless of other somatic symptoms of general distress

Illustration: GHQ-28 measurement invariance across time

- The second wave of data is available on participants, which was collected almost 10 years after the first wave
  - age = 53 at T1, age = 62 at T2
- Let's test our GHQ-28 Somatic Symptoms subscale for longitudinal invariance
  - we can do so separately for men and women since there is gender DIF



Mplus longitudinal MI model setup

```
MODEL:
Somatic1 BY T1i1-T1i4* T1i56 T1i7 (1-6); !equality of loadings
[Somatic1@0];
Somatic1@1;
[T1i1-T1i4 T1i56 T1i7]  (7-12);   !equality of intercepts
T1i1-T1i4 T1i56 T1i7  (13-18);    !equality of residuals

Somatic2 BY T2i1-T2i4* T2i56 T2i7 (1-6);
[Somatic2*];
Somatic2*;
[T2i1-T2i4 T2i56 T2i7]  (7-12);
T2i1-T2i4 T2i56 T2i7  (13-18);

Somatic2 WITH Somatic1;
T1i1-T1i4 T1i56 T1i7 PWITH T2i1-T2i4 T2i56 T2i7; !corr. residuals

OUTPUT:  MODINDICES(10);
```

Results for longitudinal MI in GHQ-28

- No large modification indices related to MI parameters were found for males
- For females, large MI were found for
  - residual of item 7 ("*hot and cold spells*")
  - loading of item 1 ("*feeling perfectly well and in good health*")
- Other interesting results
  - residuals for items 2, 4, 5, 6 and 7 were correlated over time
  - stability of the somatic factor across time was
    - corr(S1,S2) =`0.416` for males
    - corr(S1,S2) =`0.342` for females

Practical 5. Testing for MI in longitudinal SDQ data

- Strengths and Difficulties Questionnaire for a community sample (pupils year 7) administered with 1 year interval
- Testing for invariance of Externalising construct across time
- Data can be found in “SDQpupil.dat” file
- Variables have been described before
- Tasks:
  - Specify and test the fully constrained model (all parameters equal across time)
  - Test genders separately. Any observations?

How to deal with longitudinal DIF

- First any statistical findings must be interpreted by subject matter experts
- If confirmed as bias, it is advisable to either use the reduced measure or adjust for this bias in the model
- For example, one can release equality constraints in M*plus*
  a) items without DIF have item parameters equal across time points (estimated at Time 1)
  b) items with DIF have parameters estimated separately at different time points

Appropriate measures for each time point

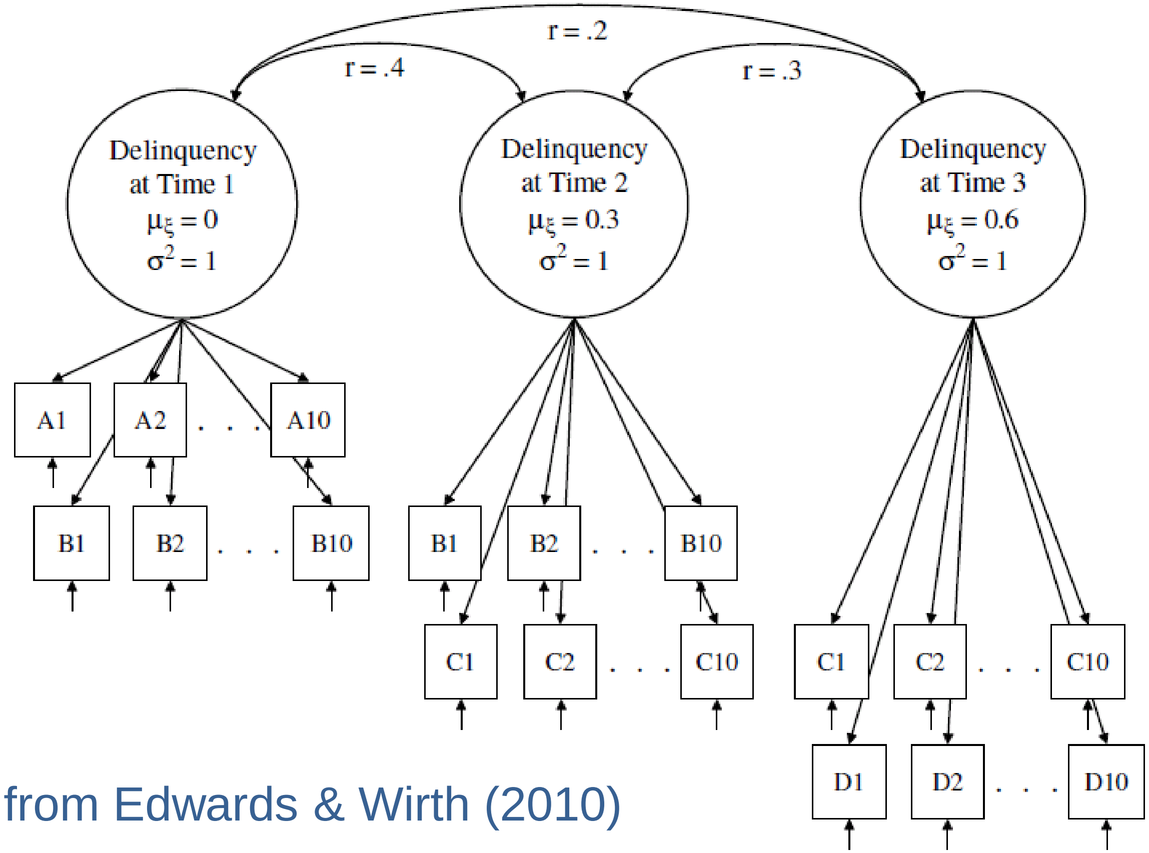

r = .2
r = .4
r = .3
Delinquency at Time 1 $\mu_\xi = 0$ $\sigma^2 = 1$
Delinquency at Time 2 $\mu_\xi = 0.3$ $\sigma^2 = 1$
Delinquency at Time 3 $\mu_\xi = 0.6$ $\sigma^2 = 1$
A1
A2
. . .
A10
B1
B2
. . .
B10
B1
B2
. . .
B10
C1
C2
. . .
C10
C1
C2
. . .
C10
D1
D2
. . .
D10

Illustration from Edwards & Wirth (2010)

SOME FINAL WORDS...

Do not mix up Predictive Invariance and Measurement Invariance

- Much of applied literature on ‘test bias’ (e.g. Jensen, 1980) were in fact referring to Predictive Invariance
- Many psychologist’s views about bias in testing are based primarily on studies that compare test/criterion regressions across populations (Hunter & Schmidt, 2000)
  - Also set in testing standards (AERA, APA & NCME, 1999; Society for Industrial/Organizational Psychology, 2003).
- Yet conclusions about test bias that rely primarily on invariance in test/criterion regressions or correlations are demonstrably flawed (Millsap, 1995; 2007).

Formal definition of PI

- Let's partition **X = (**Y, **Z**) into Y (single criterion observed score) and **Z** (a set of predictors – say a battery of selection measures)
  - *Y* might be a measure of job performance and **Z** might be a set of selection measures used to select prospective employees.

$$P(Y|\mathbf{Z},\mathbf{V}) = P(Y|\mathbf{Z}) \qquad (2)$$

  - Y = *observed score on criterion measure*
  - **Z** = *a set of measures intended to predict* Y
  - **V** = *other characteristics (often a scalar group identifier for demographic variables such as gender or ethnicity)*
- **V** *should be irrelevant to* Y *once* **Z** *is considered*

PI does not support MI

- Millsap (1995) showed that, under **realistic conditions**, prediction invariance does not support measurement invariance.
  - In fact, prediction invariance is generally indicative of *violations* of measurement invariance
  - If two groups differ in their latent means, and a test has prediction invariance across the levels of the grouping variable, it must have measurement bias with regard to group membership.
  - Conversely, when a test is measurement invariant, it will generally show differences in predictive regression parameters, when two groups differ in their latent means.

Reminder: Purposes of MI studies

- *Purpose 1: Fairness and equity in testing.*
- *Purpose 2: Dealing with a possible threat to internal validity.*
  - rule out measurement artefact 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.

Thank you

ANY QUESTIONS?

References and further reading 1

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research*: Volume 1 – The analysis of case-control studies. Lyon: International Agency for Research on Cancer.

Byrne, B.M. (1994). Testing for factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*, 289–311.

Hambleton, R.K.,Merenda, P.F. & Spielberger, C.D. (Eds.). (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale: Lawrence Erlbaum.

Jensen, A.R. (1980). *Bias in mental testing*. New York: Free Press.

Liu, I-M, & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223-1234.

Magis, D., Béland, S., Tuerlinckx, F., & Boeck, P. de (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847-862.

References and further reading 2

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127–143.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-543.

Millsap, R. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*, 461–473.

Millsap, R.E., & Kwok, O.M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*, 93–115.

Penfield, R. D. (2003). Application of the Breslow-Day test of trend in odds ratio heterogeneity to the detection of nonuniform DIF. *Alberta Journal of Educational Research, 49*, 231-243.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 44*, 187-210.

References and further reading 3

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti Estimator of the Cumulative Common Odds Ratio to DIF Detection in Polytomous Items. *Journal of Educational Measurement, 40*, 353-370.

Penfield, R. D., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*, 295-312.

Scheuneman, J.D. (1982). A new look at bias in aptitude tests. In P. Merrifield (Ed.), *New directions for testing and measurement: Measuring human abilities, No. 12*. San Francisco: Jossey-Bass.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159–194.

Stark, S., Chernyshenko, O., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292-1306.

Thurstone, L.L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

References and further reading 4

Vijver, F.J.R. van de, & Poortinga, Y.H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*(1), 29-37.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods,* 3, pp. 4-70.

Wu, A.D., Li, Z. & Zumbo, B.D. (2007). Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data. *Practical Assessment, Research & Evaluation, 12*, 1-26.

Zumbo, B. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language assessment Quarterly, 4(2)*, 223–233.

Zwick, R. (1990). When do item response function and Mantel–Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185–197.