Lecture 12: Differential Item Functioning (DIF)

DIF analyses require that the test-taking population can be split into two or more groups, based on gender or ethnicity, for example. Then an item displaying DIF is an item whose relationship to the latent trait is different for the two groups. These two groups are sometimes called the *majority* and *minority* groups, or else the *reference* and *focal* groups.

For example, if we are comparing the performances of males and females on a math test, we might find that for a given item, a higher difficulty parameter exists for males than females. This would imply that for males and females of the same ability levels (θs) males will have a lower probability of correct response.

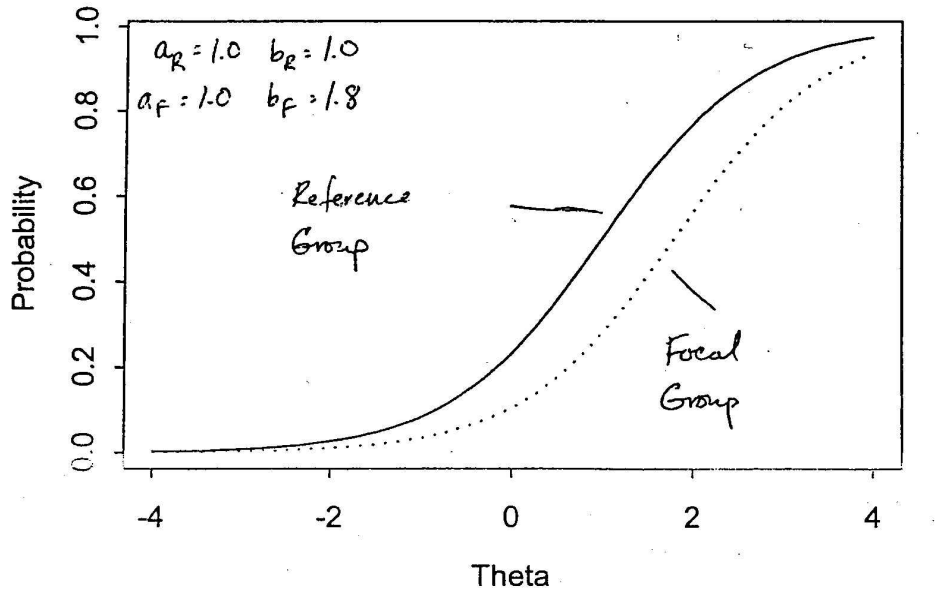
Central to any DIF analysis is the identification of a set of items that can be assumed to perform the same across groups. These items, called *anchor items*, define a common ability metric against with the remaining items can be tested for DIF.

Definitions of DIF

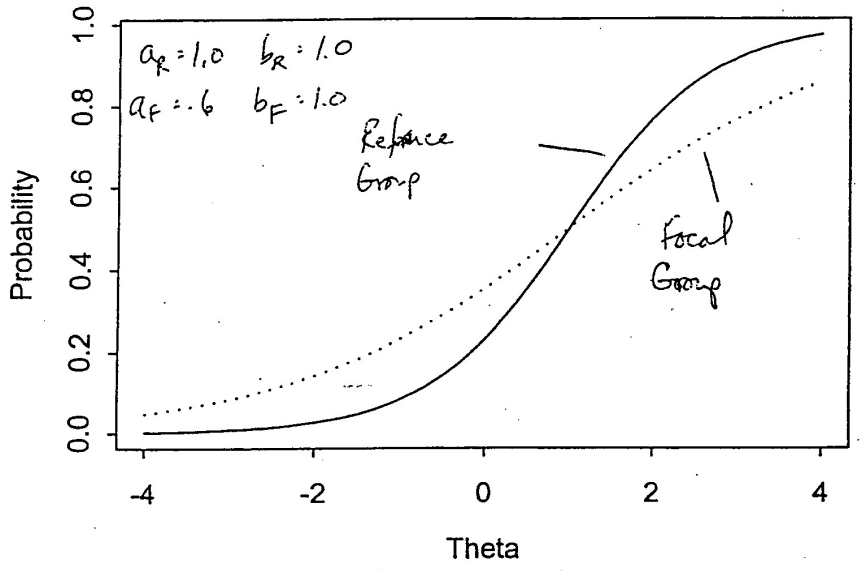
- CTT: An item shows DIF if the reference and focal groups differ in their mean performance on the item (i.e., the *p*-values are different for the two groups)

- IRT: An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right.

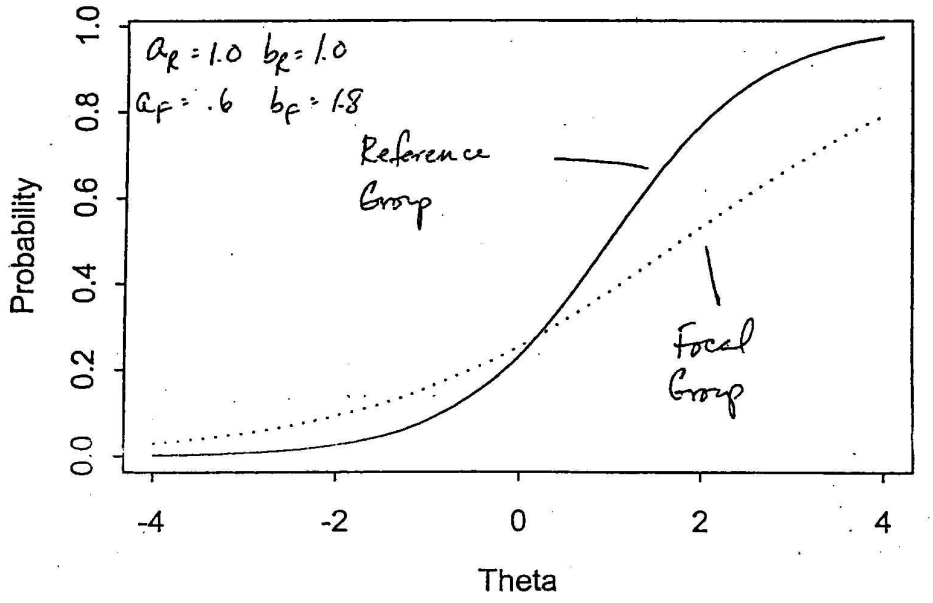
Example of a 2PL item demonstrating DIF with respect to difficulty parameter:



Example of a 2PL item demonstrating DIF with respect to discrimination parameter:



Example of a 2PL item demonstrating DIF with respect to both difficulty and discrimination parameter:



In order to perform a DIF analysis, separate calibrations of item parameters must be obtained for the two groups. Once the calibrations are done, there is still a need to link the parameter estimates across groups so that they are on the same ability scale. After the item parameter estimates are on the same scale, there are several procedures that can be used to test for DIF.

Testing for DIF

- χ^2 test (Lord, 1980): Tests for difference in the item parameters across the two groups.

$H_0 : b_R = b_F ; a_R = a_F ; c_R = c_F$

H_A : At least one of parameters varies across groups.

$$\chi^2 = (a_{diff} b_{diff} c_{diff})' \sum^{-1} (a_{diff} b_{diff} c_{diff})$$

where $a_{diff} = a_R - a_F ; \quad b_{diff} = b_R - b_F ; \quad c_{diff} = c_R - c_F$

$\sum^{-1} =$ the variance-covariance matrix of the difference between the item parameter estimates

Assuming the sample size is large, the test statistic is a χ^2 with a p degrees of freedom, where p is the number of parameters being compared.

Some limitation of this approach: 1) differences in item parameters do not necessarily imply meaningful differences in the ICCs, 2) the test statistic tends to have a high false-positive rate.

- Area measures: These measures are based on the estimated areas between ICCs for the two groups being compared.

To estimate this area, the θ scale is divided into discrete intervals, and the difference between ICCs is estimated within each interval. Then an estimate of the area between ICCs is computed as:

$$A_i = \sum_{\theta} \left| P_R(\theta) - P_F(\theta) \right| \Delta\theta$$

Raju (1988, 1990) computed an exact formula for the area between ICCs in the 1PL, 2PL, and 3PL models. For the 3PL (assuming the guessing parameters is the same across groups), the area is:

$$(1-c) \left| \frac{2(a_R - a_F)}{D a_R a_F} \log[1 + e^{D a_R a_F (b_R - b_F)/(a_R - a_F)}] - (b_R - b_F) \right|$$

A test statistic can be derived using formulae for the standard error of the area measure. Otherwise, empirical cut values can be derived for determining if the area represents statistically significant DIF.

- Likelihood Ratio (LR) test (Thissen, Steinberg and Wainer, 1988): Compares the log-likelihood when an item's parameters are constrained to be the same across groups versus to the log-likelihood when the parameters are free to vary. Provided a set of anchor items has

been identified, the LR tests for DIF in a studied item compares two models - a compact model and an augmented model.

Data: PCL-R (Psychopathy Checklist Revised) items for Caucasian males versus. African-American males

```
33222222112132121311                    1
12222322211221231211                    1
22311223213333221322                    1
11311233311123321111                    1
23233323211121331223                    1
                    233333232 2 323211112
                    32 333333 313332 1312
                    123122122332223123122
                    2132321233323221 3332
                    2133223223222321 3122
                    122223232331232311132
                    2233333233132311 3112
```

Example of Compact model:

```
DIF for individual items
>PRO RA IN NG=2 NI=40 NE=3052;
>TEST ALL GR NC=(3(0)40);
>EQUAL AJ IT=(21(1)40) WI=(1(1)20);
>EQUAL BK=(1,2) IT=(21(1)40) WI=(1(1)20);
>EST NC=1000;
>END;
3
123
1111111111111111111111111111111111111
2222222222222222222222222222222222222
3333333333333333333333333333333333333
(T41, I1, T1, 40A1)
```

Example of Augmented Model:

```
DIF for individual items
>PRO RA IN NG=2 NI=40 NE=3052;
>TEST ALL GR NC=(3(0)40);
>EQUAL AJ IT=(22(1)40) WI=(2(1)20);
>EQUAL BK=(1,2) IT=(22(1)40) WI=(2(1)20);
>EST NC=1000;
>END;
3
123
1111111111111111111111111111111111111
2222222222222222222222222222222222222
3333333333333333333333333333333333333
(T41, I1, T1, 40A1)
```

Results for Item 1 (Glibness/superficial charm)

```
0TOTAL,    NEGATIVE TWICE THE LOGLIKELIHOOD,      ALL GROUPS=  61238.7
0TOTAL,    NEGATIVE TWICE THE LOGLIKELIHOOD,      ALL GROUPS=  61211.8
```

- Nonparametric methods: These methods do not use IRT model-fitting to evaluate the occurrence of DIF.

  1) Mantel-Haenzsel (MH) procedure (Holland & Thayer, 1988): Tests whether the odds of correct versus incorrect responses differ across groups. A conditional odds ratio is defined as

$$\alpha = [R_{rm} / W_{rm}] / [R_{fm} / W_{fm}]$$

where R_{rm} and R_{fm} are the number of right answers on the item for examinees in the reference and focal groups having total test score m respectively, and W_{rm} and W_{fm} are the numbers of wrong answers on the item for examinees in the reference and focal groups having total test score m respectively.

A chi-square statistic based on these odds ratios reduces to

$$MH - \chi^2 = \left[\left| \sum_m R_{rm} - \sum_m E(R_{rm}) \right| - .5 \right]^2 / \sum_m Var(R_{rm}),$$

where

$$E(R_{rm}) = N_{rm}(R_{rm} + R_{fm}) / N_m$$

and

$$Var(R_{rm}) = [N_{rm}(R_{rm} + R_{fm}) N_{fm}(W_{rm} + W_{fm})] / [N_m^2 (N_m - 1)]$$

This statistic is distributed as a χ^2 with 1 degree of freedom under a null hypothesis that the common odds ratio is 1. To quantify the amount of DIF in an item, a constant odds ratio can be computed

$$\alpha_{MH} = \left[\sum_m R_{rm} W_{fm} / N_m \right] / \left[\sum_m R_{fm} W_{rm} / N_m \right].$$

This index is then transformed to produce

$$MHD - DIF = -2.34 \ln[\alpha_{MH}].$$

DIF STATISTICS: DICHOTOMOUS ITEMS

Name	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
Var 2	0.011	-0.052	0.233	-0.223	0.034	OK	A
Var 3	1.786	0.358	0.246	1.455	0	OK	A
Var 4	0.189	0.128	0.23	0.557	2.776	OK	A
Var 5	0	-0.031	0.236	-0.131	0.03	OK	A
Var 6	0.161	-0.13	0.25	-0.52	0.043	OK	A
Var 7	0.323	-0.162	0.238	-0.681	1.156	OK	A
Var 8	0.001	0.036	0.238	0.151	2.109	OK	A
Var 9	0.013	0.053	0.229	0.231	1.11	OK	A
Var 10	0.001	-0.025	0.252	-0.099	0.016	OK	A
Var 11	0.42	-0.194	0.25	-0.776	0.202	OK	A

DIF STATISTICS: POLYTOMOUS ITEMS

Name	Mantel	L-A LOR	LOR SE	LOR Z
Var 2	94.397	0.948	0.1	9.48
Var 3	0.061	0.023	0.093	0.247
Var 4	0.508	-0.067	0.094	-0.713
Var 5	0.748	-0.079	0.093	-0.849
Var 6	4.55	-0.2	0.094	-2.128
Var 7	1.61	-0.119	0.093	-1.28
Var 8	1.428	-0.112	0.093	-1.204
Var 9	1.501	-0.114	0.094	-1.213
Var 10	0.543	-0.069	0.093	-0.742
Var 11	2.797	-0.156	0.093	-1.677

2) Standardization procedure (Dorans and Kullick, 1986) / SIBTEST & Poly-SIBTEST (Shealy & Stout, 1993; Roussos & Stout, 1996)

Estimates the expected item score difference across groups conditional on ability level. To test for DIF in a studied item, the user must first specify a valid subtest of items (often all of the remaining items). The valid subtest score serves as a surrogate for an ability level. Conditional on each valid subtest score level (k), the average score on the studied item is computed for both the reference and focal groups - $\overline{Y}_{rk}$ and $\overline{Y}_{fk}$. And these values are adjusted to new values $\overline{Y}^{*}_{rk}$ and $\overline{Y}^{*}_{fk}$ that represent the average scores for members of each group at the same ability level.

The DIF index β_{UNI} is computed as:

$$\hat{\beta}_{UNI} = \sum_{k}[\overline{Y}^{*}_{rk} - \overline{Y}^{*}_{fk}]g(k)$$

where $g(k)$ is the proportion of examinees having valid subtest score k .

$\hat{\beta}_{UNI}$ is asymptotically normally distributed under a null hypothesis of no DIF, and thus can be tested for statistical significance.

DIF-Pack (SIBTEST, Polu-SIBTEST, am Crossing SIBTEST for items and bundles of items) is available at https://psychometrics.onlinehelp.measuredprogress.org/tools/dif/ for free download.

*** Readings on DIF: http://www.rcmar.ucla.edu/content/irt-dif