# Sofia University
# Saint Kliment Ohridski

## DEPARTMENT FOR LANGUAGE TEACHING AND INTERNATIONAL STUDENTS

## ALTE Member

# A Differential Item Functioning Study for Less Widely-Taught Languages

**Authors:**

**Mardik Andonyan, Ph.D.**

**Julia Todorinova**

# The study targeted:

- items that might show *bias* against some significant groups of test-takers
- further improvement of Standardized Test in Bulgarian at B2 level

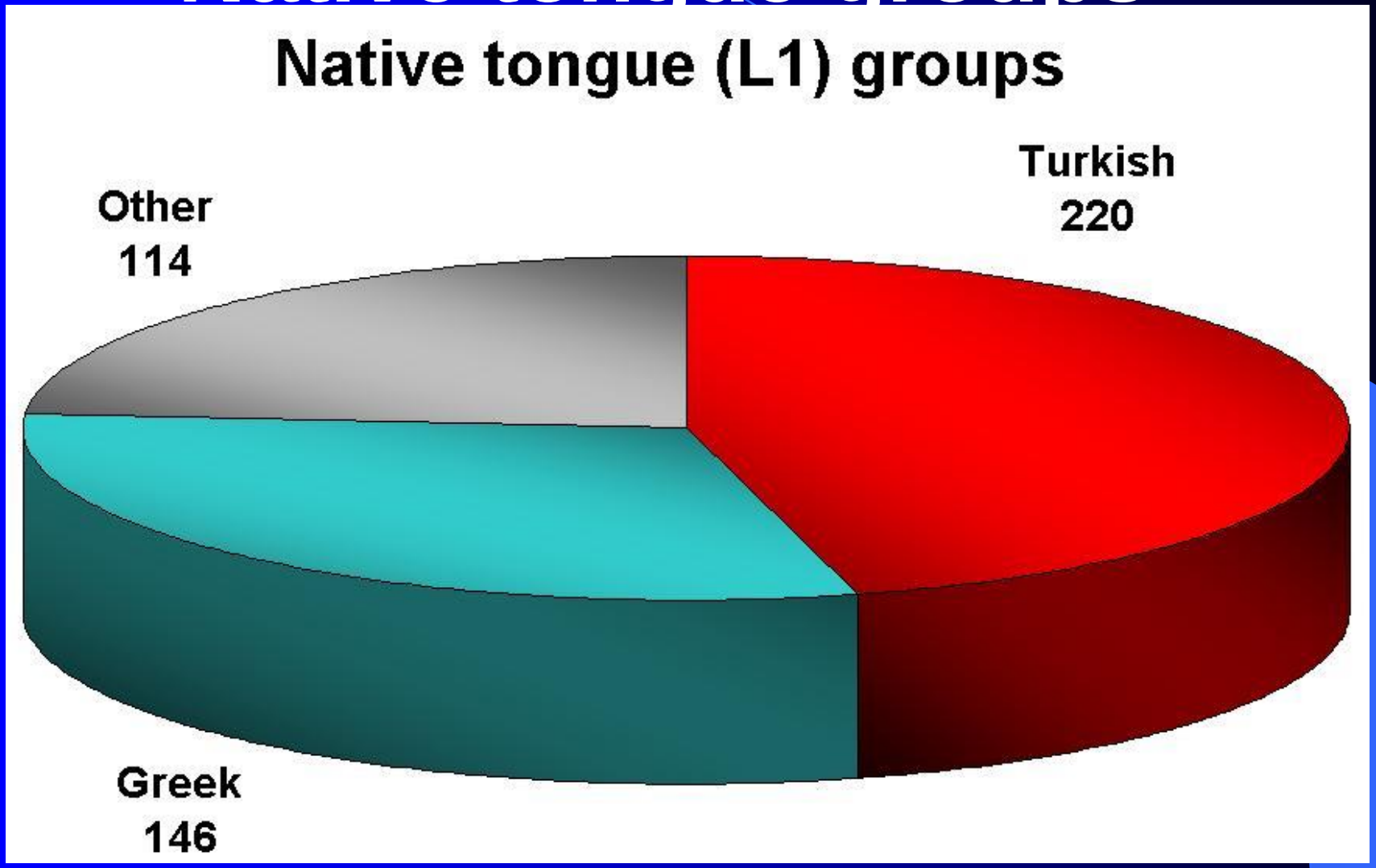
**The Tested population consisted of 480 test-takers distributed as follows:**

- **345 foreign students at the Department for Language Teaching**
- **100 individual test-takers at the Department**
- **35 individual test-takers at our examination centre in Thessaloniki, Greece.**

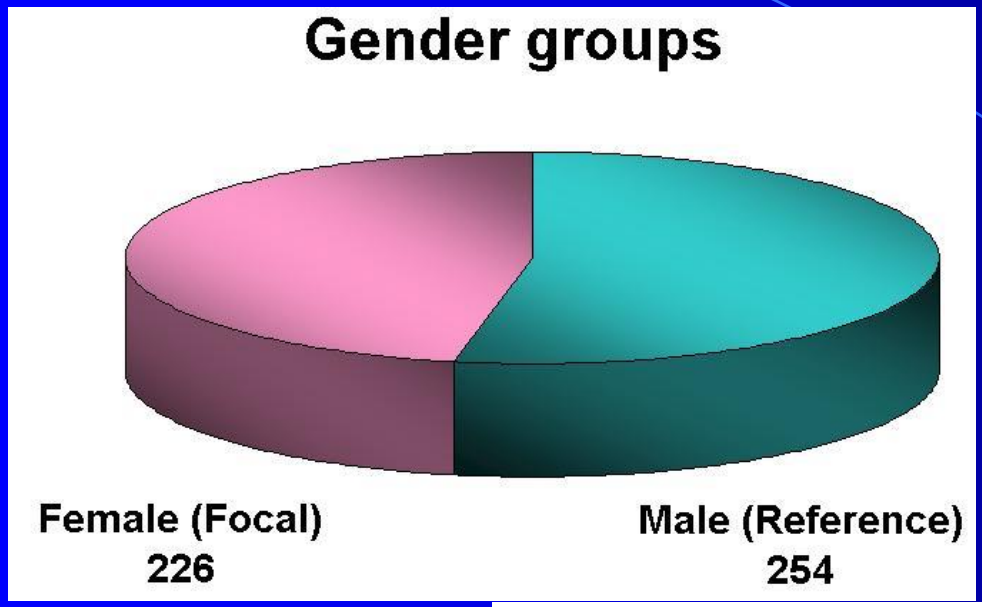
## Significant Groups of examinees were formed in regard to:

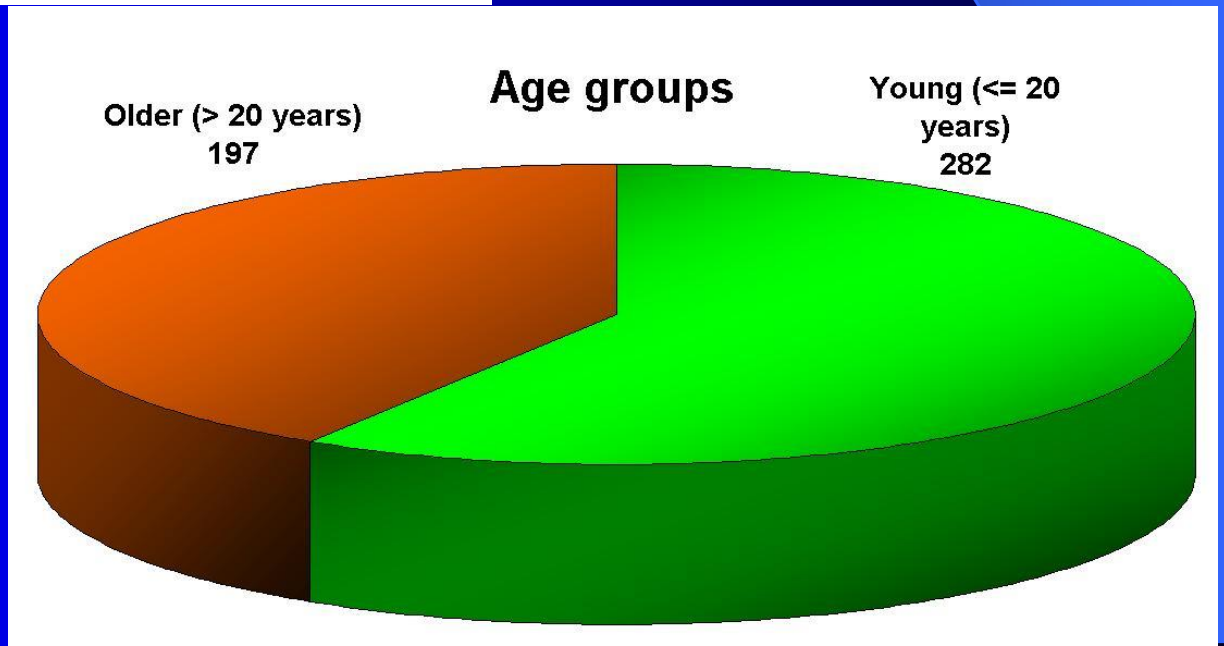
- Native tongue
- Gender
- Age
- Education

# Native tongue groups

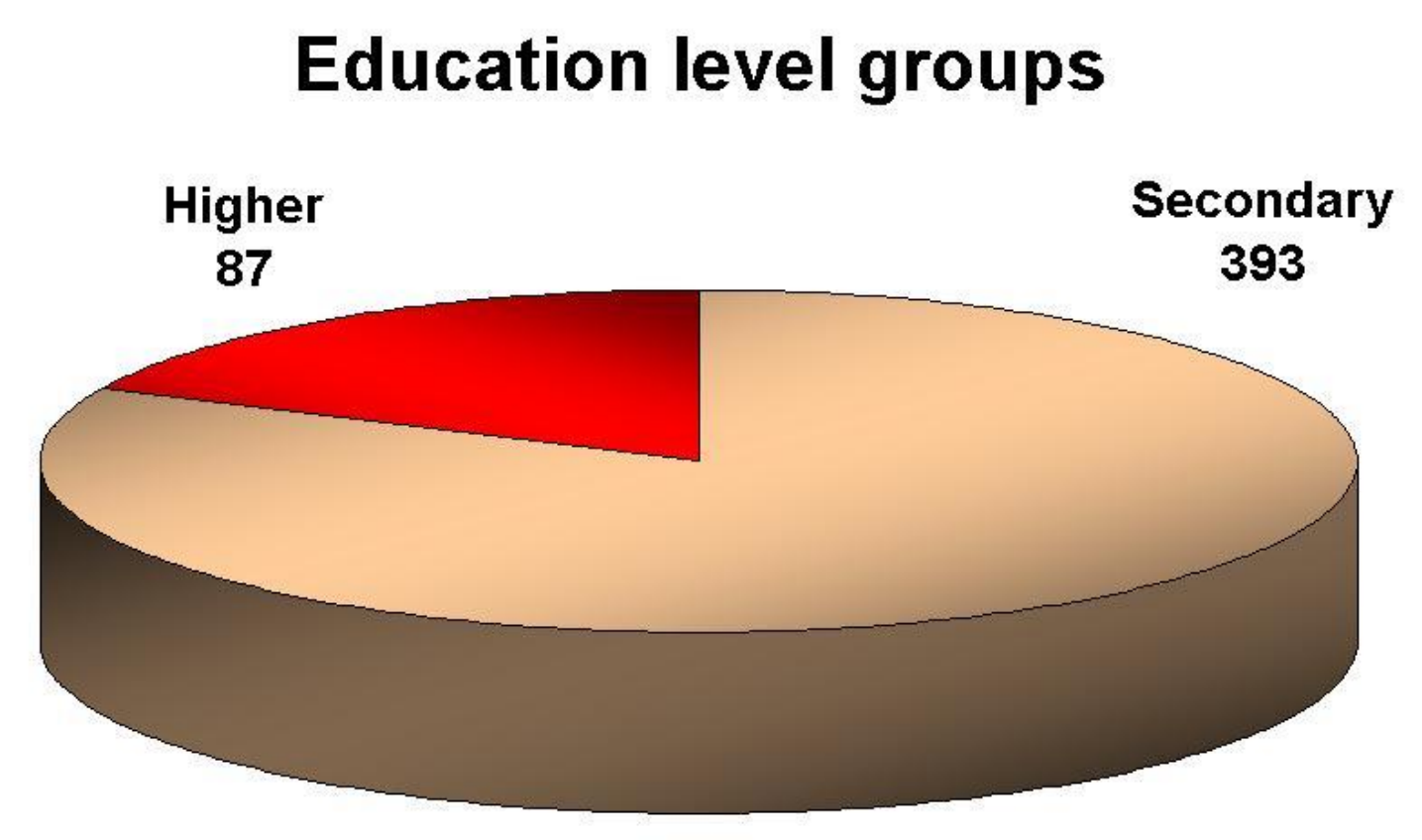


# Gender and Age Groups



# Education Level Groups



# DIF detecting methodology used:

- The nonparametric MH procedure of Mantel and Haenszel, proposed by Holland and Thayer in 1988
- MH Δ-DIF statistic compares the odds of a correct response to an item for the focal group to that of the reference group.

# DIF Detecting Software

- Most of DIF detecting computer programmes are designed to operate with huge amount of input data.
- They calculate the odds of a correct response to an item for the focal and reference groups over each test score point.
- They do not work with relatively small samples of input data.

# Overcoming the problem

**EZDIF computer programme written by Niels G. Waller was used to analyze the uniform and non-uniform DIF because it handles problems of virtually any size.**

# Overcoming the problem

**EZDIF computer programme provides user with control over conditioning-level bin widths, which is very important for small samples that are not large enough to cover each score point with necessary representatives from both reference and focal groups.**

# Ability Levels Defined

**Five ability levels were defined for the purpose of our DIF study as follows:**

| Ability level | F | D | C | B | A |
|---|---|---|---|---|---|
| Score range | 0<br>12 | 13<br>23 | 24<br>29 | 30<br>37 | 38<br>45 |

# EZDIF software features

**EZDIF measures DIF in two ways:**

**a) with the Mantel-Haenszel (Holland and Thayer, 1988) procedure the uniform DIF is detected and measured;**

**b) with the Logistic Regression (Narayanan & Swaminathan, 1996) procedure non-uniform DIF can be detected.**

# EZDIF software features

- It allows using the real test item labels.
- It analyzes DIF in a two-stage manner so that items showing large DIF in the first stage are automatically removed from the matching variable in the second stage.
- It is a completely free software.

# ETS (Educational Testing Service) DIF classification

- **A- level or *negligible* DIF**
- **B- level or *moderate DIF***
- **C-level or *large* DIF**

# EZDIF software output

- common odds ratio $\alpha$
- Mantel-Haenszel chi-square statistics $\chi^2$ and its significance level
- MH $\Delta$-DIF statistic and its standard error
- ETS DIF size code A, B or C
- Empirical Item Characteristic Curves
- Logistic Regression output

# EZDIF output for Mantel-Haenszel procedure

**Screenshot:**

Results for Pass Number: 1

| ETS | ITEM | Alpha | X^2 | P-Value | MHD-DIF | SE (MHD-DIF) |
|---|---|---|---|---|---|---|
| A | 201 | 1.160 | 0.071 | 0.790 | 0.348 | 0.799 |
| A | 202 | 1.141 | 0.051 | 0.821 | 0.309 | 0.786 |
| A | 203 | 0.938 | 0.009 | 0.925 | 0.149 | 0.645 |
| A | 204 | 0.619 | 2.482 | 0.115 | 1.129 | 0.661 |
| A | 205 | 1.285 | 0.443 | 0.506 | -0.589 | 0.717 |
| A | 206 | 0.930 | 0.007 | 0.936 | 0.170 | 0.725 |
| A | 207 | 1.124 | 0.047 | 0.828 | -0.274 | 0.739 |

# Interpreting EZDIF output for Mantel- Haenszel procedure

- The MH technique is very simple, easy to implement, does not require large sample sizes and also provides statistics that have tests of significance.
- Size effect of uniform DIF is easily detected by ETS codes – A, B or C.
- However, it is not powerful in detecting nonuniform DIF.

# The output for Swaminathan and Rogers Logistic Regression procedure

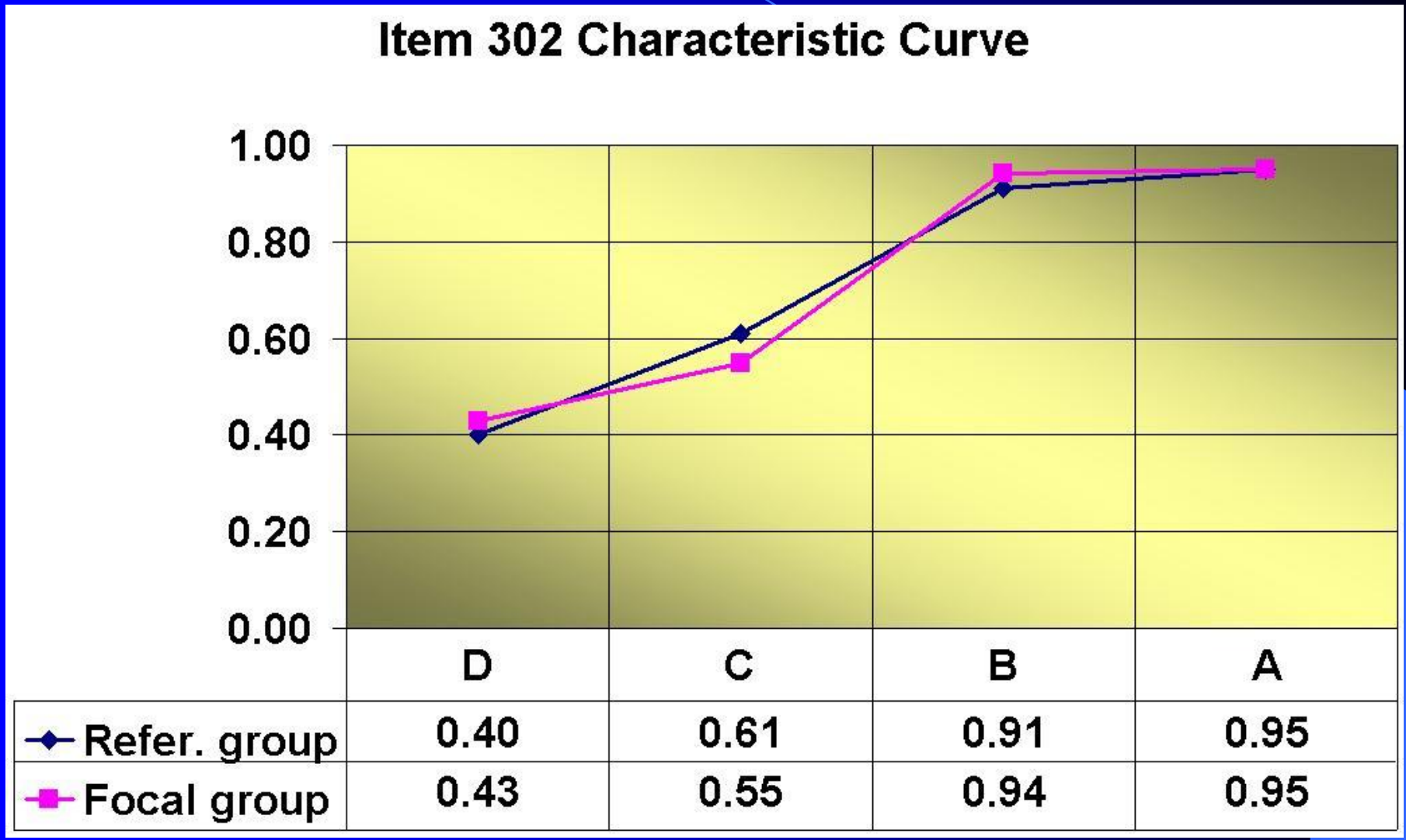
**Screenshot:**

| Item Number:209 | Estimate | SE Estim | Z | p-value |
|---|---|---|---|---|
| Intercept | -1.871244 | 0.483211 | -3.872519 | 0.000115 |
| Trait | 0.110279 | 0.013925 | 7.919373 | 0.000000 |
| Group | 0.060784 | 0.881146 | 0.068983 | 0.945017 |
| Trait x Group | -0.007219 | 0.027850 | -0.259212 | 0.795525 |

# The output for Swaminathan and Rogers Logistic Regression procedure

- An item exhibits **uniform DIF** if the *Group* statistics is different from 0, and ***Trait x Group* statistics is 0**.
- If *Trait x Group* statistics is different from 0, then **nonuniform DIF** is present irrespective of the *Group* statistics.

# Item Characteristic Curves

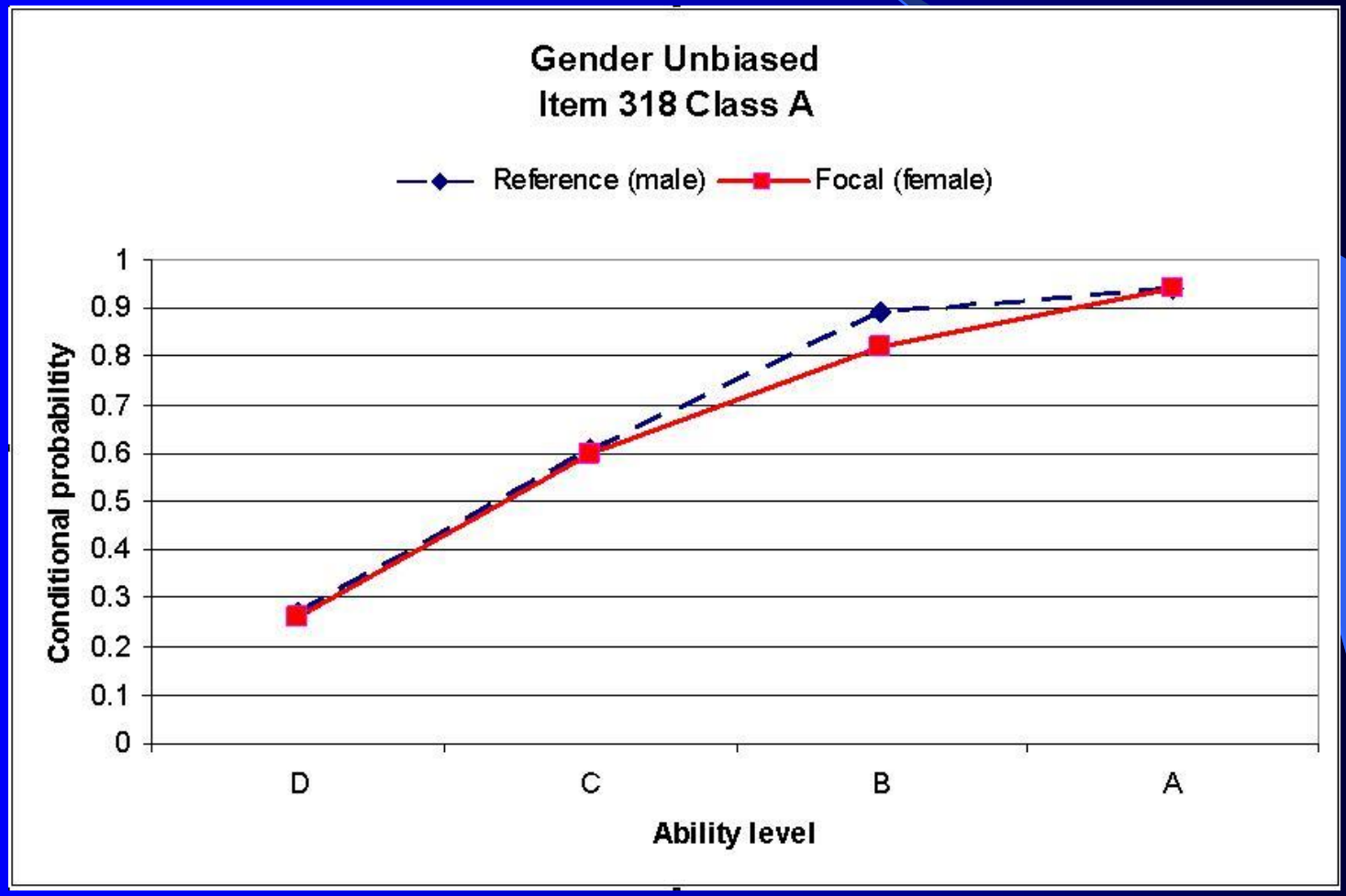


# Mantel-Haenszel Gender DIF Results

| Level of DIF | Number of items | List of Items | Flagged Items | Removed Items |
|---|---|---|---|---|
| Negligible (A-level) | 42 | | | |
| Moderate (B level) | 3 | 223, 307, 312 | 223, 307, 312 | |
| Large (C-level) | none | | | none |

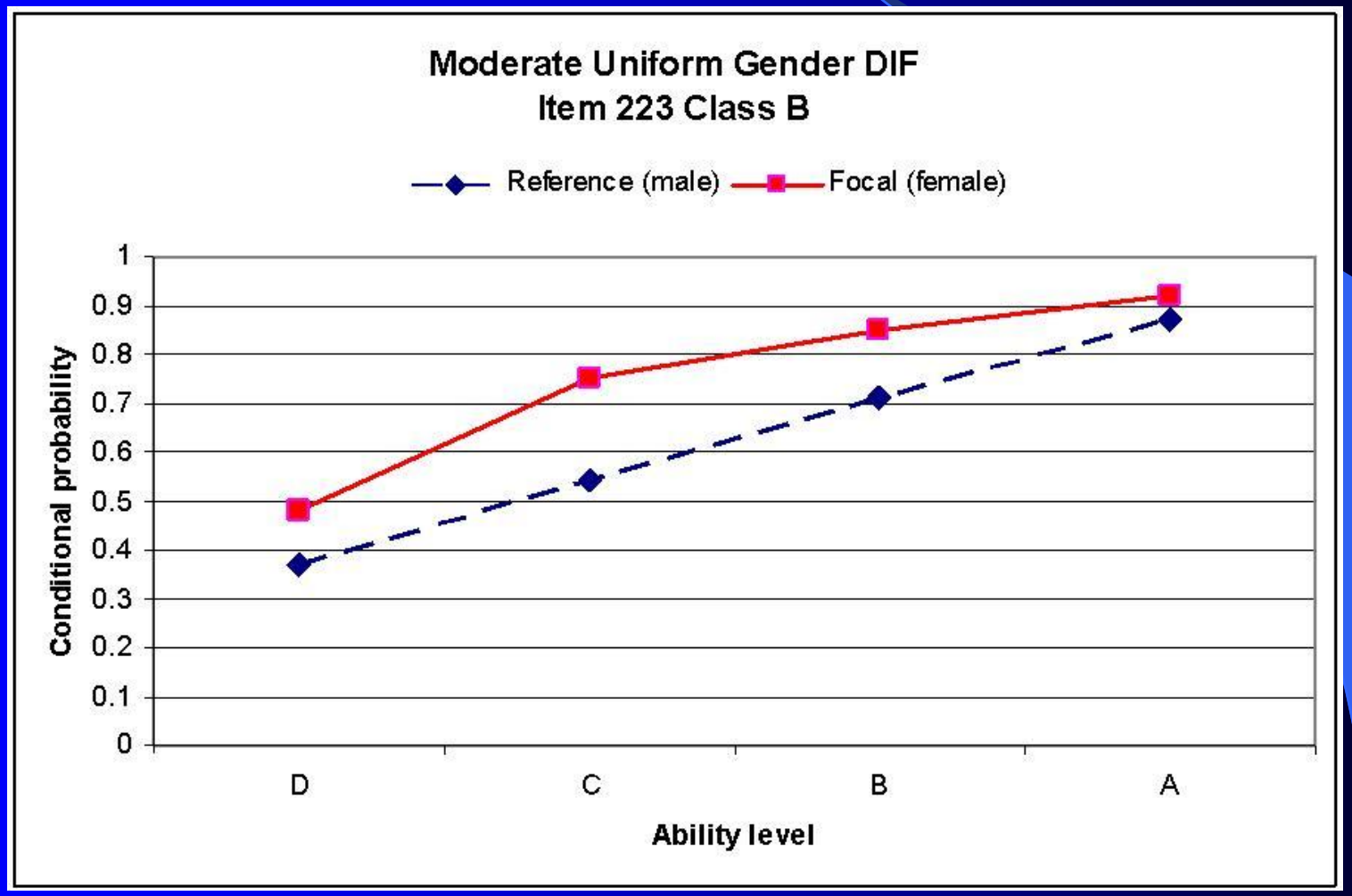
# Sample Item Curves

## (a) Gender Unbiased Item



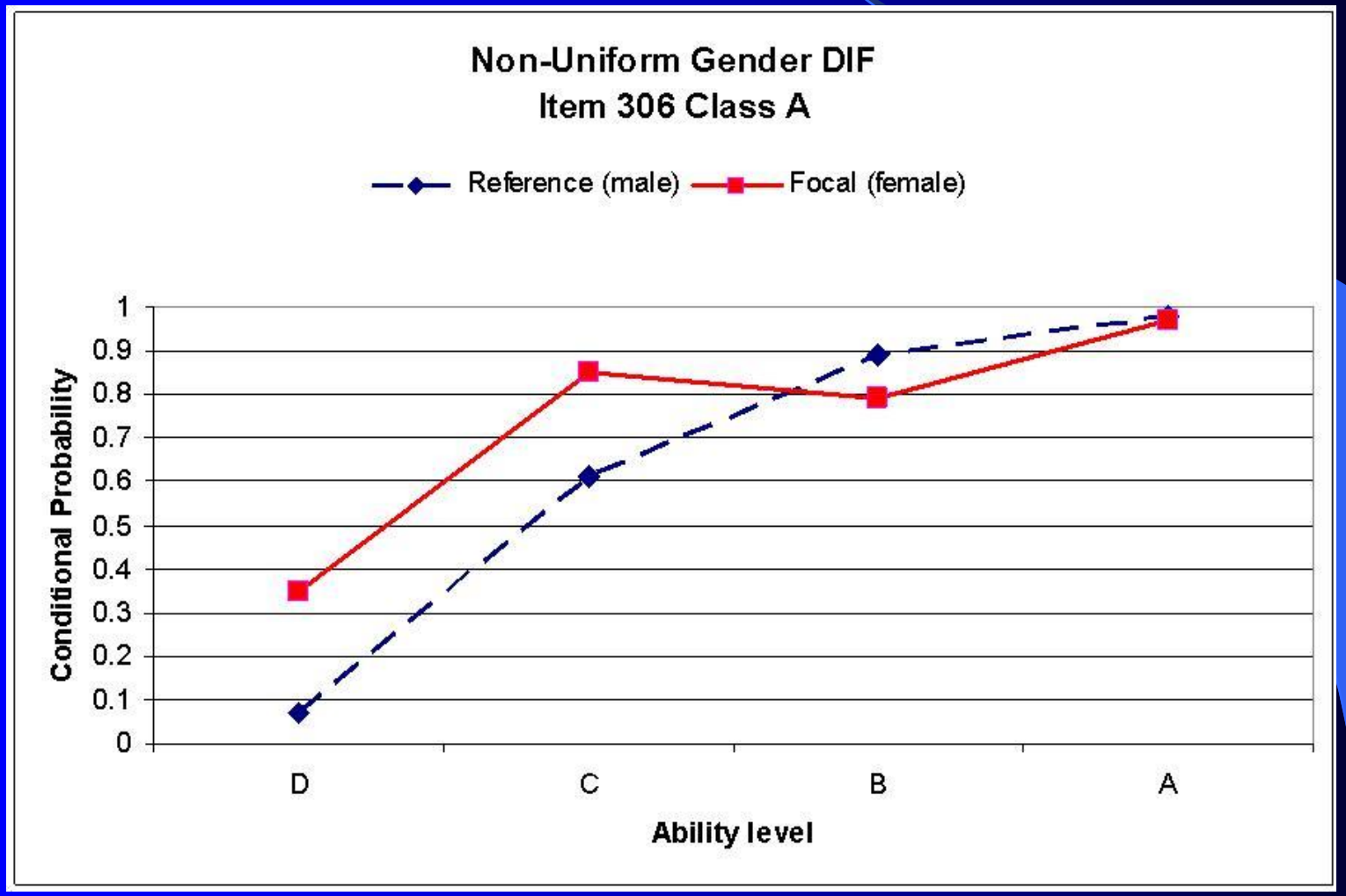
# Sample Item Curves

## (b) Flagged Item (Moderate Uniform DIF)



# Sample Item Curves

## (c) Item with negligible nonuniform bias



# DIF analysis against native tongue (L1) bias

| Group name | Absolute size | Percentage |
|---|---|---|
| Turkish | 220 | 45.84% |
| Greek | 146 | 30.41% |
| Other | 114 | 23.75% |

# Mantel-Haenszel First Pass
## native tongue (L1) DIF results

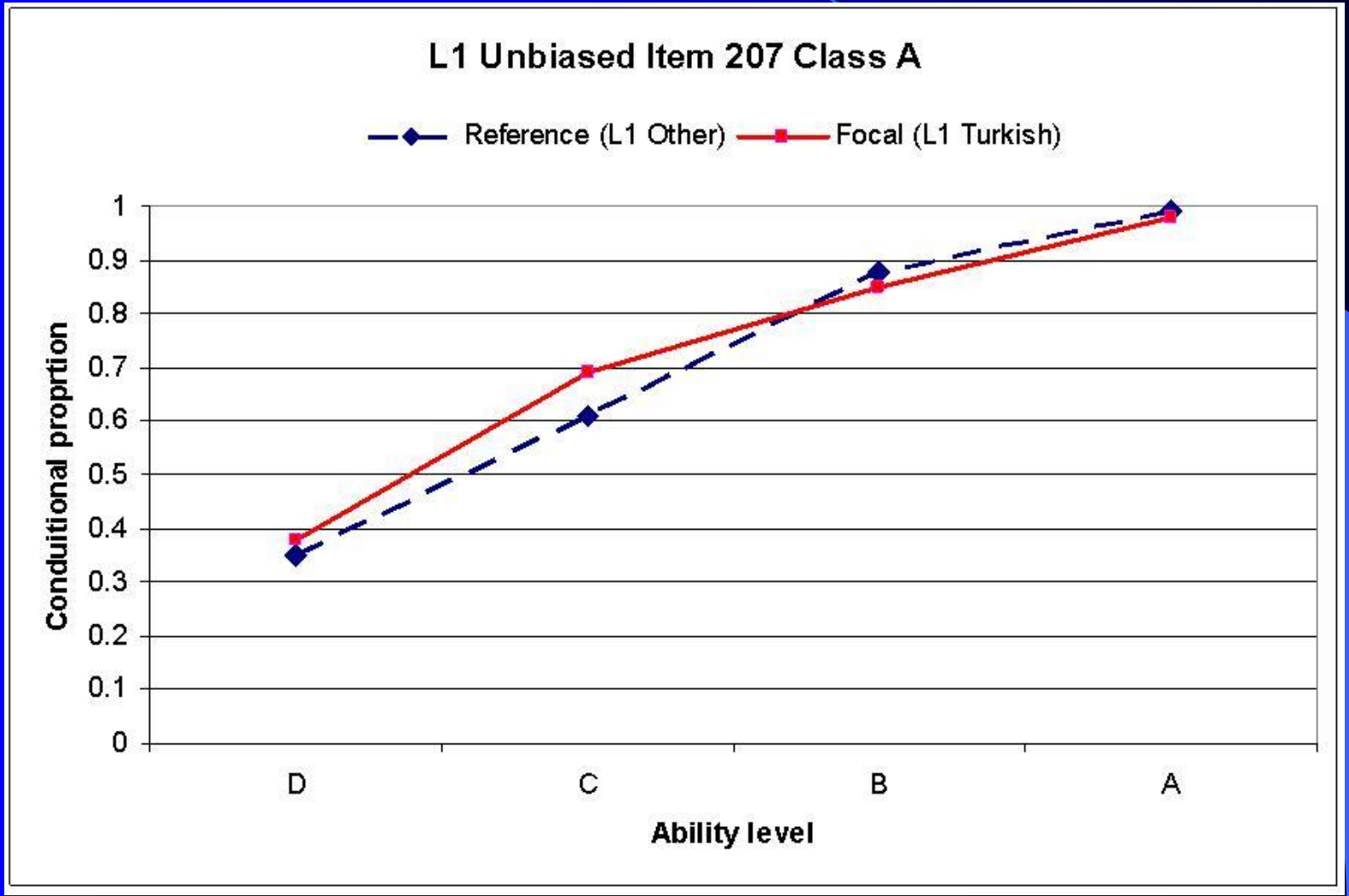
| Level of DIF | Number of items | List of Items | Flagged Items | Removed Items |
|---|---|---|---|---|
| Negligible (A-level) | 39 | | | |
| Moderate (B level) | 5 | 213,214, 224, 316, 320 | 213, 214, 316, 320 | 224 |
| Large (C-level) | 1 | 318 | | 318 |

# Mantel-Haenszel DIF Results Second pass after removing items 224 and 318

| Level of DIF | Number of items | List of Items | Flagged Items | Removed Items |
|---|---|---|---|---|
| Negligible (A-level) | 38 | | | |
| Moderate (B level) | 5 | 213,214, 301, 313, 320 | 320 | 224 |
| Large (C-level) | 0 | 0 | | 318 |

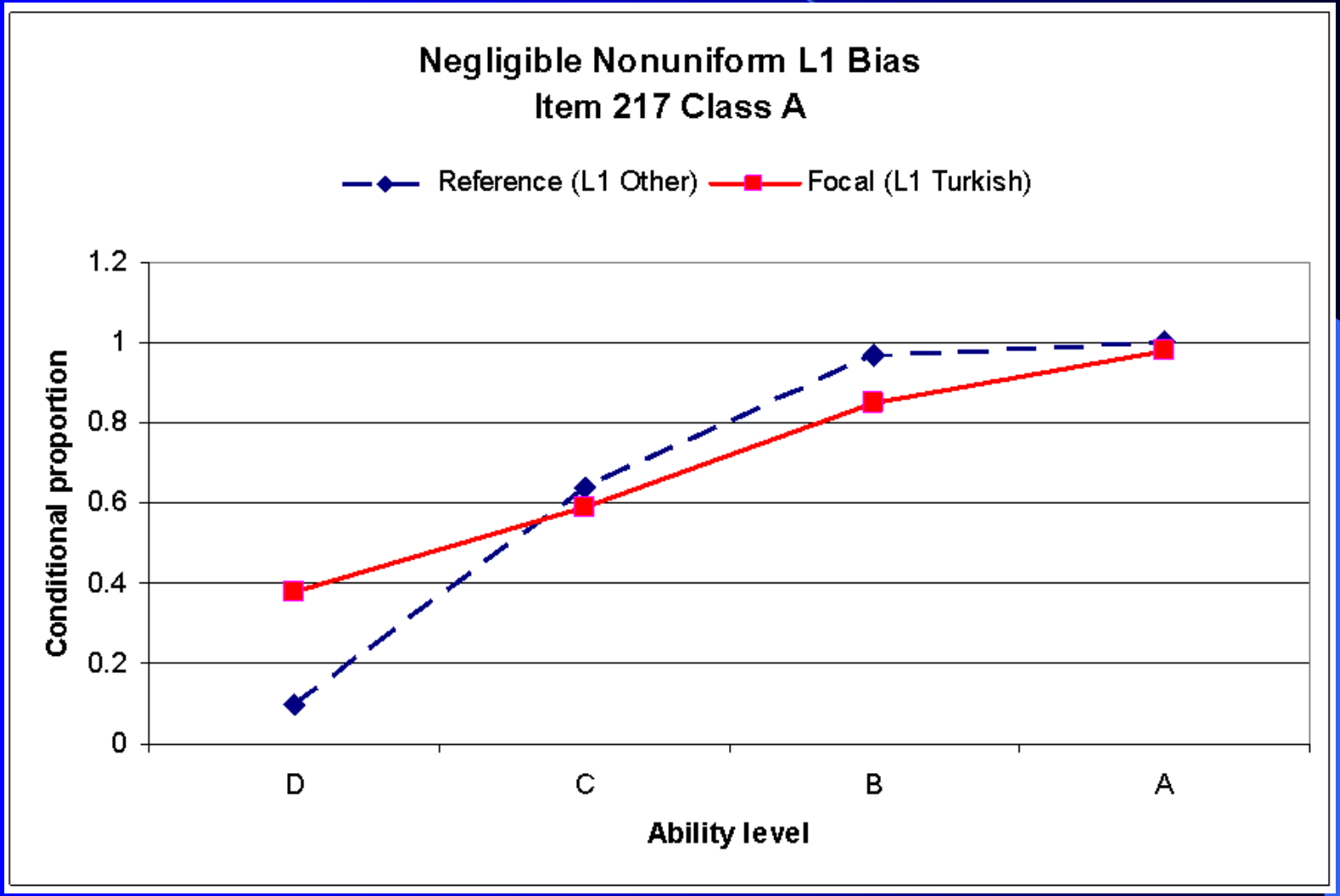
# Typical Item Curves for L1 bias

## (a) L1 Unbiased Item



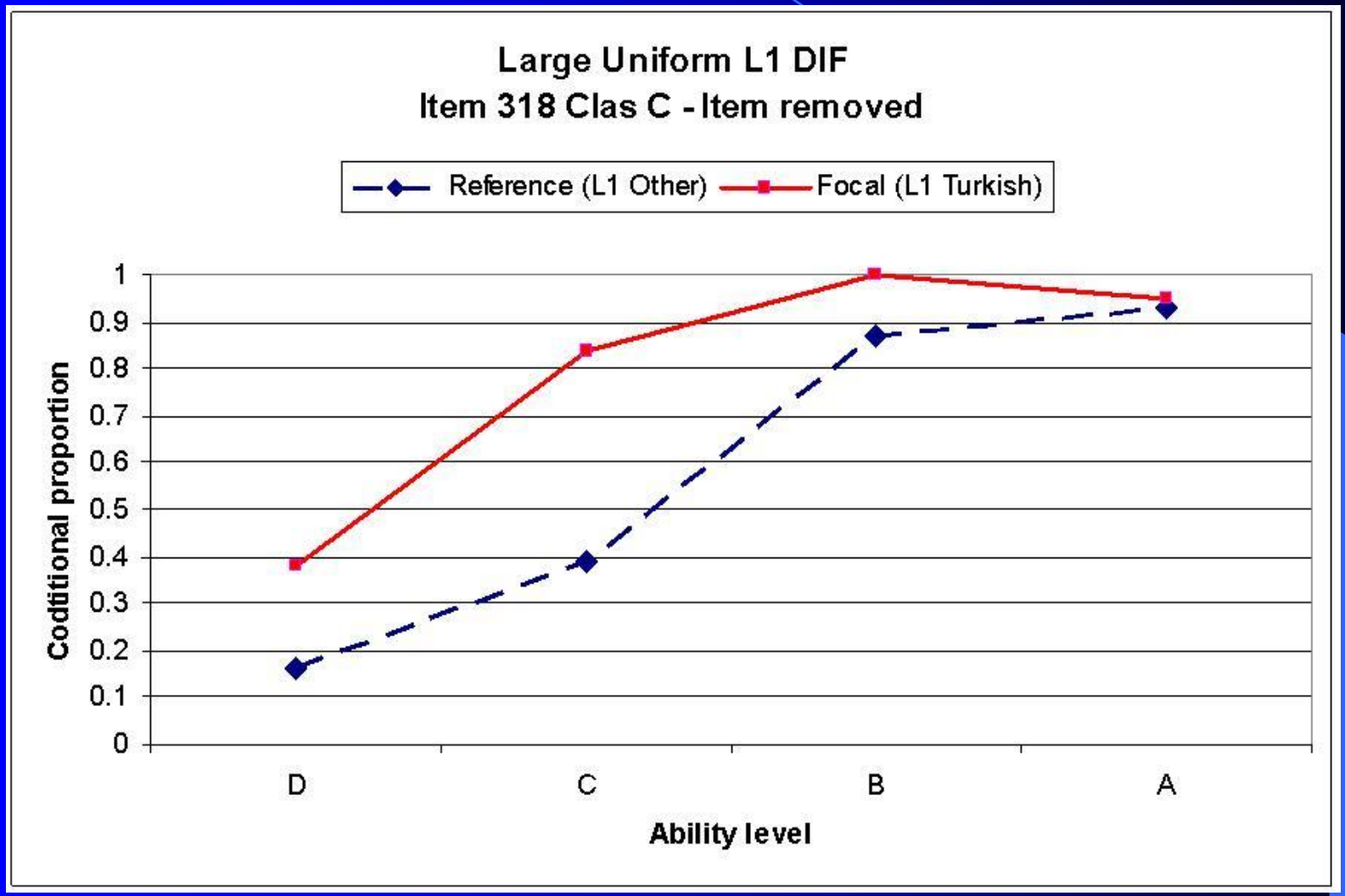
# Typical Item Curves for L1 bias

## (b) Item with negligible nonuniform L1 bias



# Typical Item Curves for L1 bias

## (c) Removed item with large uniform L1 bias



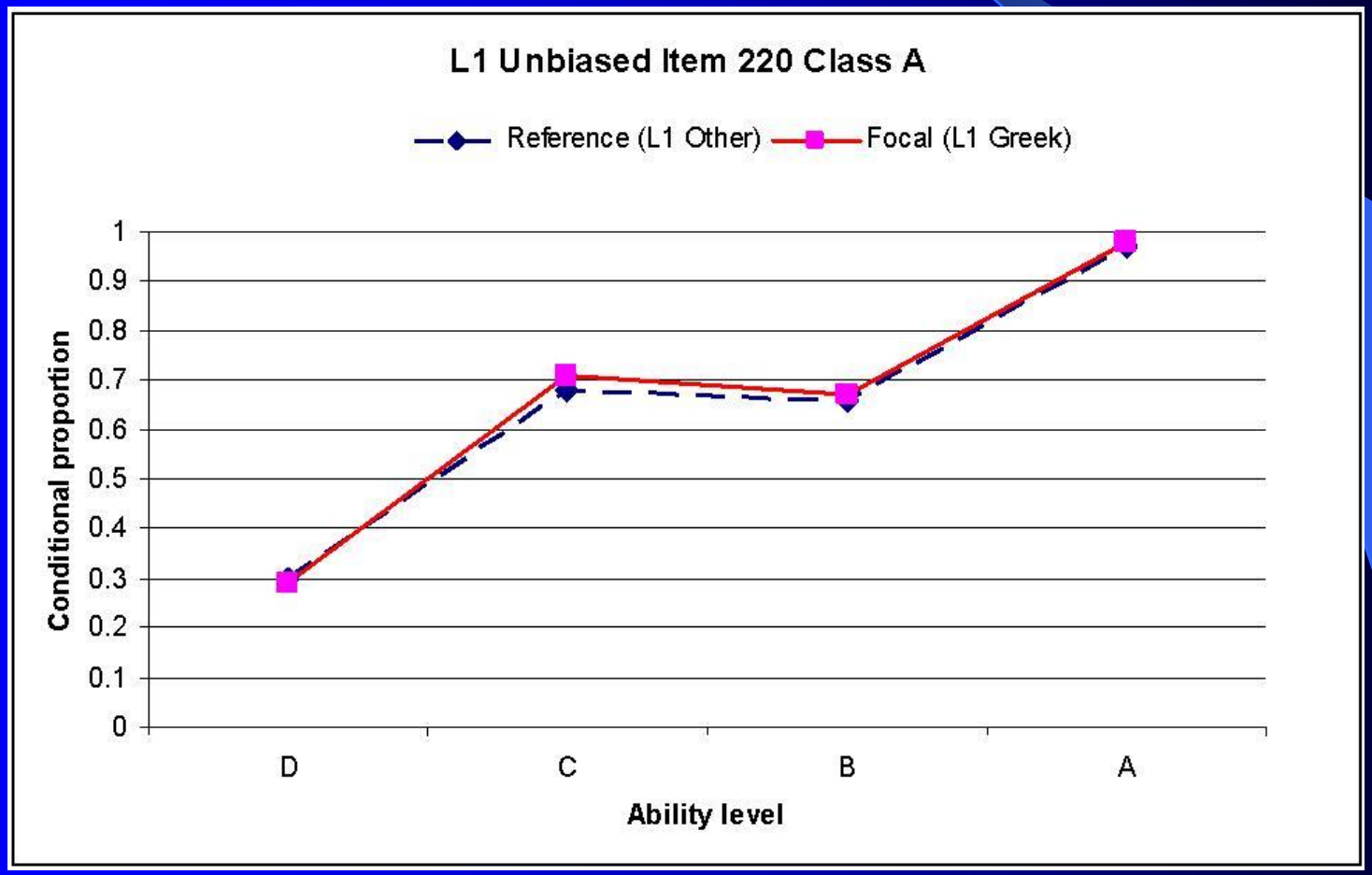
# L1 DIF analyses
# Greek versus Others
## Mantel-Haenszel DIF Results

| Level of DIF | Number of items | List of Items | Flagged Items | Removed Items |
|---|---|---|---|---|
| Negligible (A-level) | 41 | | | |
| Moderate (B level) | 4 | 203,213, 219, 224 | 213 | 224 - already removed |
| Large (C-level) | none | | | none |

# L1 DIF analyses

## Greek versus Others

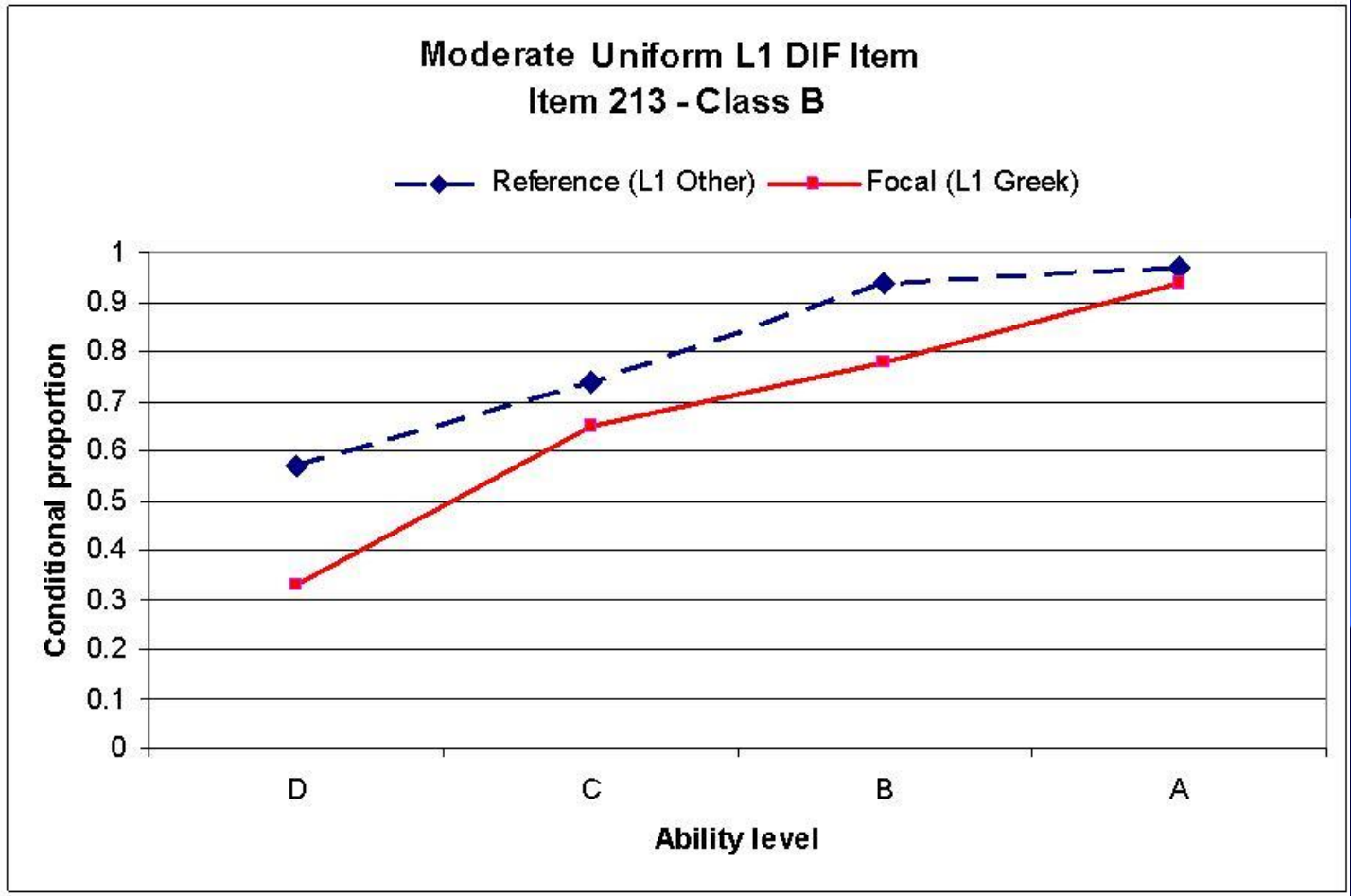
### (a) Unbiased item



# L1 DIF analyses

## Greek versus Others

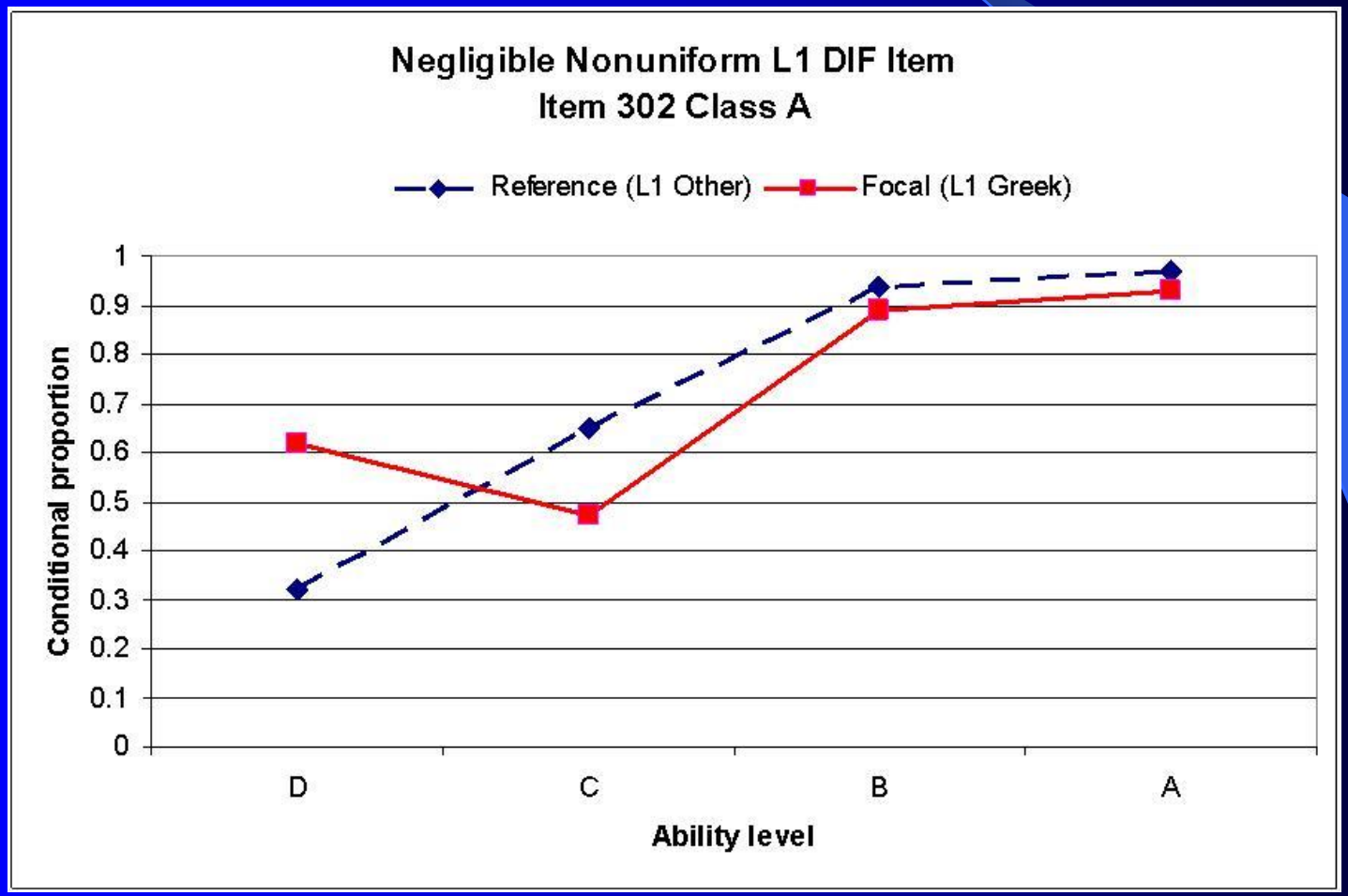
### (b) Item with moderate uniform DIF



# L1 DIF analyses

## Greek versus Others

### (c) Item with moderate nonuniform DIF



# L1 DIF analyses

## Greek versus Turkish

### Mantel-Haenszel DIF Results

| Level of DIF | Number of items | List of Items | Flagged Items | Removed Items |
|---|---|---|---|---|
| Negligible (A-level) | 39 | | | |
| Moderate (B level) | 5 | 213, 219, 316, 318,320 | 213 | |
| Large (C-level) | 1 | 224 | | 224 - already removed |

# DIF analysis against age bias

## Examined population samples

| Group name | Size | Percentage |
|---|---|---|
| Young (<= 20 years) | 282 | 58.75% |
| Older (> 20 years) | 197 | 41.25 |

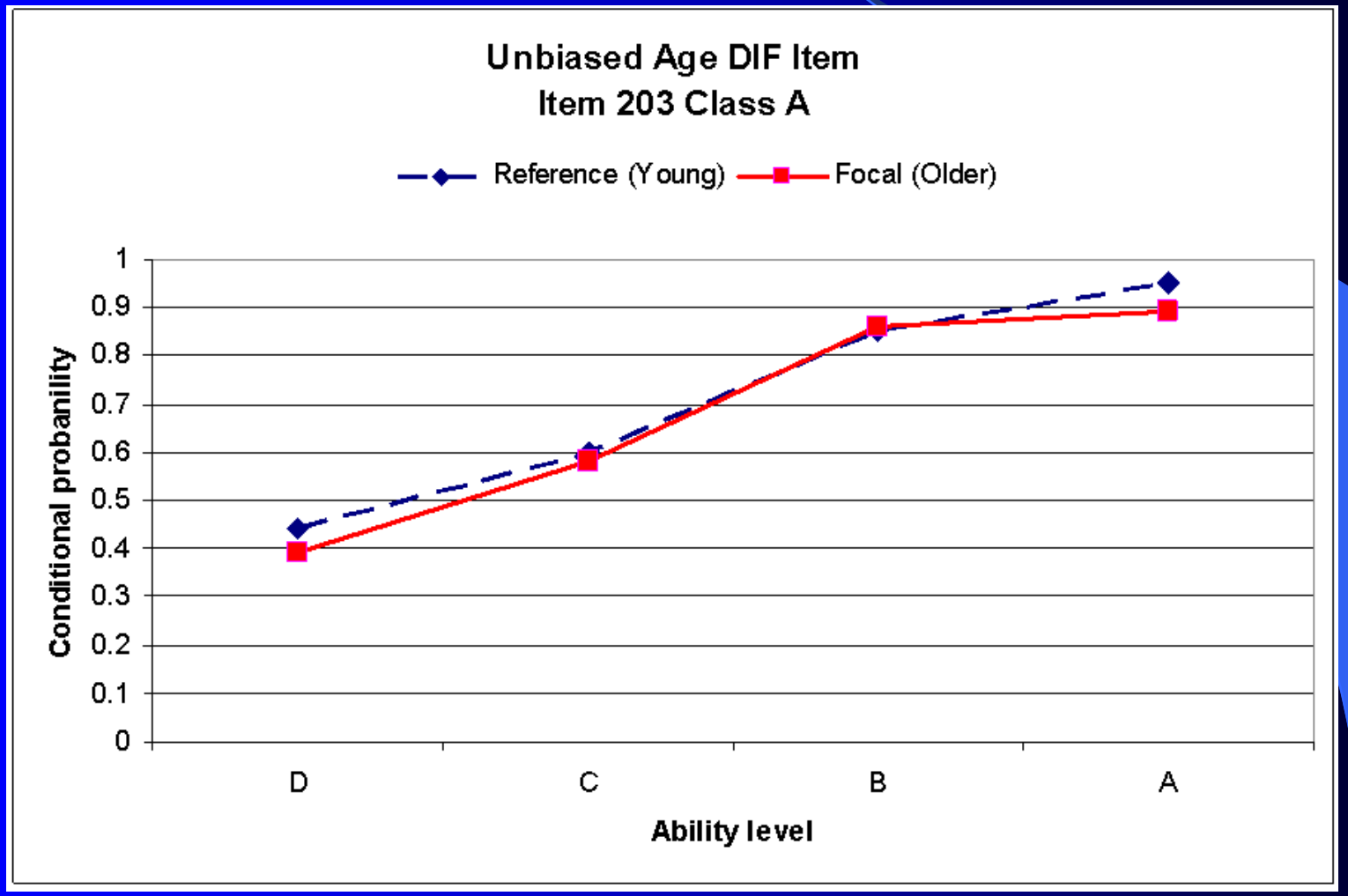
# DIF analysis against age bias

## Mantel-Haenszel age DIF Results

| Level of DIF | Number of items | List of Items | Flagged Items | Removed Items |
|---|---|---|---|---|
| Negligible (A-level) | 35 | | | |
| Moderate (B level) | 10 | 201,204, 301, 304, 310, 319, 320 | 304 | |
| Large (C-level) | none | | | |

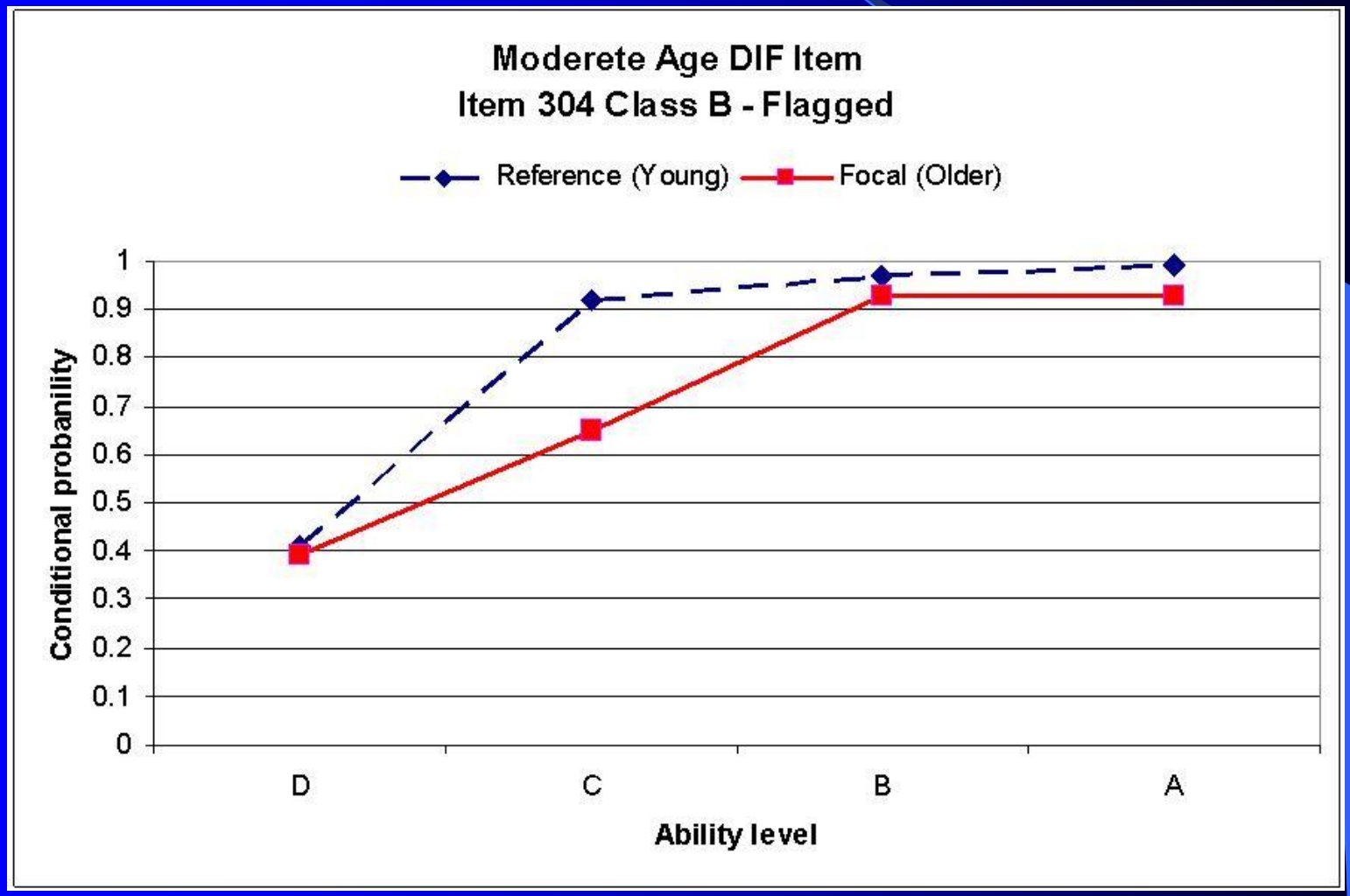
# DIF analysis against age bias

## (a) Unbiased item



# DIF analysis against age bias

## (b) Item with moderate uniform age DIF



# DIF analysis against education level bias

## Examined population samples

| Group name | Size | Percentage |
|---|---|---|
| Secondary | 393 | 81.88% |
| Higher | 87 | 18.12% |

# DIF analysis against education level bias

## Mantel-Haenszel DIF Results

| Level of DIF | Number of items | List of Items | Flagged Items | Removed Items |
|---|---|---|---|---|
| Negligible (A-level) | 41 | | | |
| Moderate (B level) | 3 | 201, 308, 310 | 201, 310 | |
| Large (C-level) | 1 | 320 | | 320 – already removed |

# Mantel-Haenszel DIF Results
## Summary Table

| Level of DIF | Number of items | List of Items | Flagged Items | Removed Items |
|---|---|---|---|---|
| Bias Free or Negligible (A-level) | 27 | 202, 205, 206,. 207, 208, 209, 210, 211, 212, 215, 216, 217, 218, 220, 221, 222, 225, 302, 303, 305,306, 309, 311, 314, 315, 316, 317, | | |
| Moderate (B level) | 16 | 201, 203, 204, 213, 214, 219, 223, 301, 304, 307, 308, 310, 312, 313, 319, 320 | 6 item as follows: 213, 219, 301, 304, 310, 320 | 1 item - 320 |
| Large (C-level) | 2 | 224, 318 | | 2 items – 224 and 318 |

# Mantel-Haenszel DIF Results Summary

(1) 27 out of 45 items ( 60%) are bias free or demonstrate negligible bias – ETS code A

(2) 16 out of 45 items (35,5%) demonstrate moderate bias – EST code B.

(3) 2 out of 45 items (4,5%) demonstrate large bias – EST code C.

# Mantel-Haenszel DIF Results Summary

(4) 6 out of 45 items (13%), which have code B and appear in more than one bias list are flagged for further investigation.

(5) 2 out of 45 items (4,5%) demonstrating large bias – ETS code C, are removed from the test.

# Conclusions

**1. The conducted DIF analyses were directed to the most significant groups presented in the tested population.**

# Conclusions

**2. The investigation revealed no or very small amount of DIF against the gender, age and education level .**

**This is a fact of great importance for us, because almost all of our individual test-takers fall into these groups.**

# Conclusions

**3. Some negligible to moderate DIF was detected against the candidates having Turkish as L1. In fact, this was the largest group of students, taught at the Department, and they did consist 46% of the tested population.**

# Conclusions

**This does not mean at all that the test items are flawed, but we do suggest that part of the problems are due to a combination of factors such as discipline, motivation to work hard, attendance in language classes, background, culture, etc.**

# Conclusions

**4. We suggest that more information below the mean of the test score distribution would be desirable in future. This might be accomplished by substituting the easiest test items with items capable to enhance measurement precision for candidates tending to score much over the mean of the test score distribution.**

# Thank you
# for your attention

**Paris, April 2014**