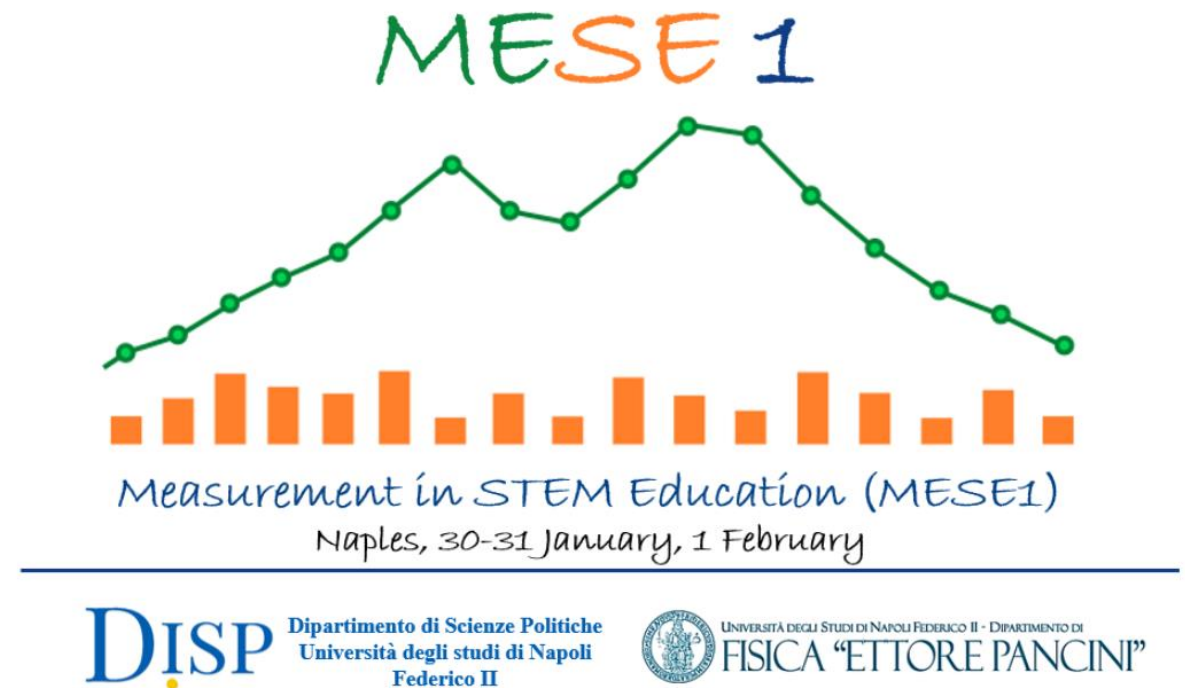


An application of
Differential Item and
Bundle Functioning
analysis to the study of
gender differences in
Mathematics Education:
implications for
educational practitioners.

Clelia Cascella

clelia.cascella@invalsi.it



Research aims and scopes

- Most of empirical research about gender differences employed DIF analysis, but
 - quantitative analysis stops with the calculation of DIF; and
 - differences relative to single item can be small (negligible) and often not statistically significant.
- The aim of the current investigation is to show how Differential Bundle Functioning can be used to interpret (gender) differences in mathematics.

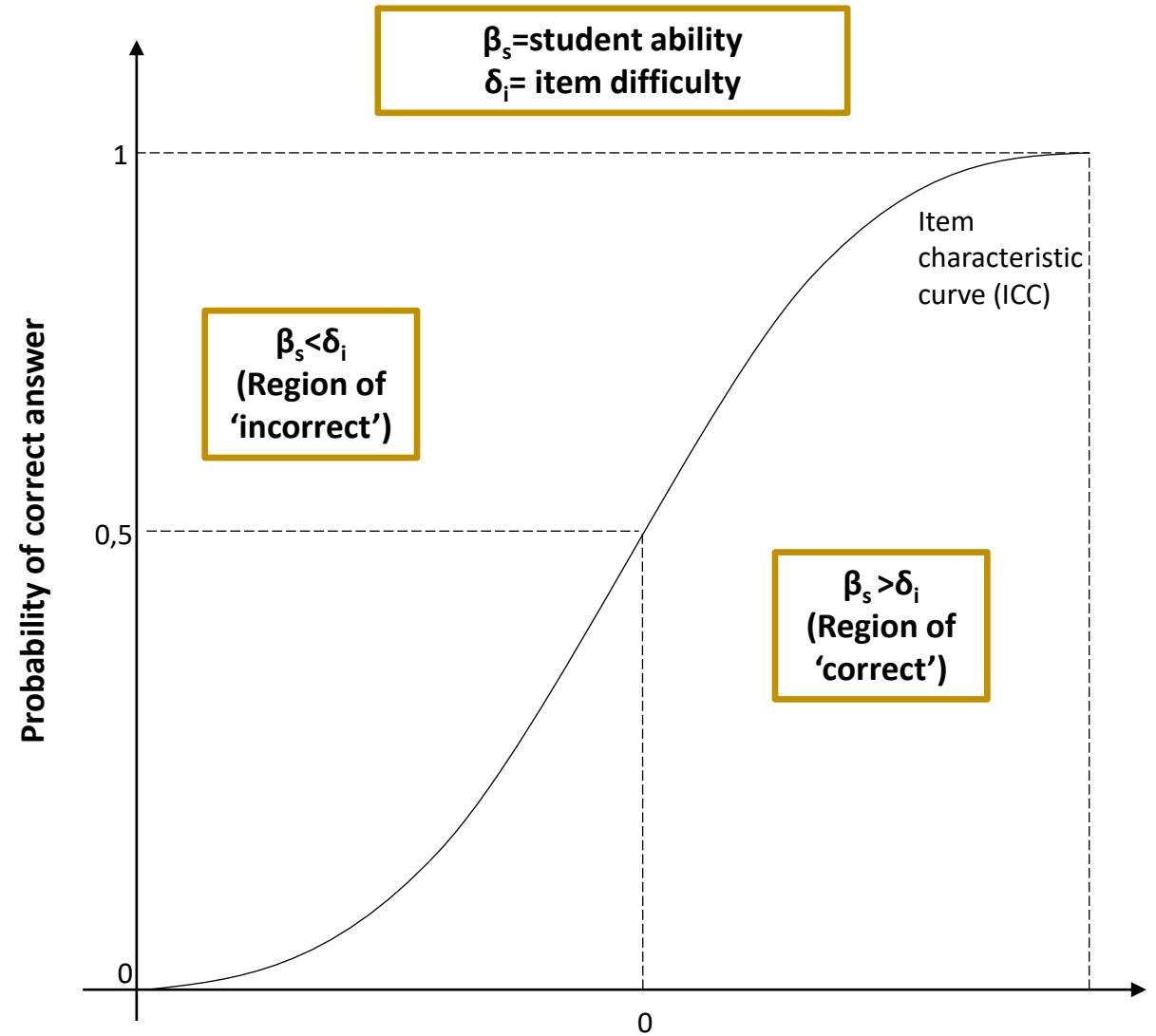



The Rasch model

- The probability of encountering an item correctly depends on student's **relative** ability, i.e. his/her ability (β_s) compared with item's difficulty (δ_i):

$$P(Y_{si} = 1 | \beta_s, \delta_i) = \frac{\exp\{\beta_s - \delta_i\}}{1 + \exp\{\beta_s - \delta_i\}}$$

- Y_{si} is the answer given by student s to the item i , with $y_{si} \in [0,1]$ and $\delta_i \in \mathbb{R}$.



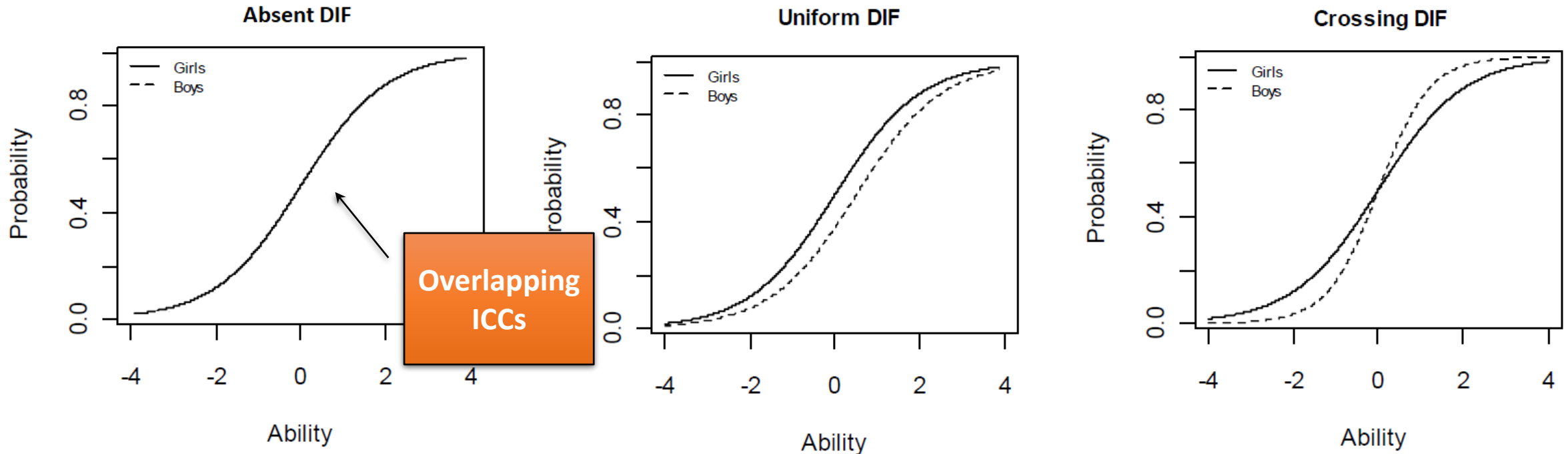


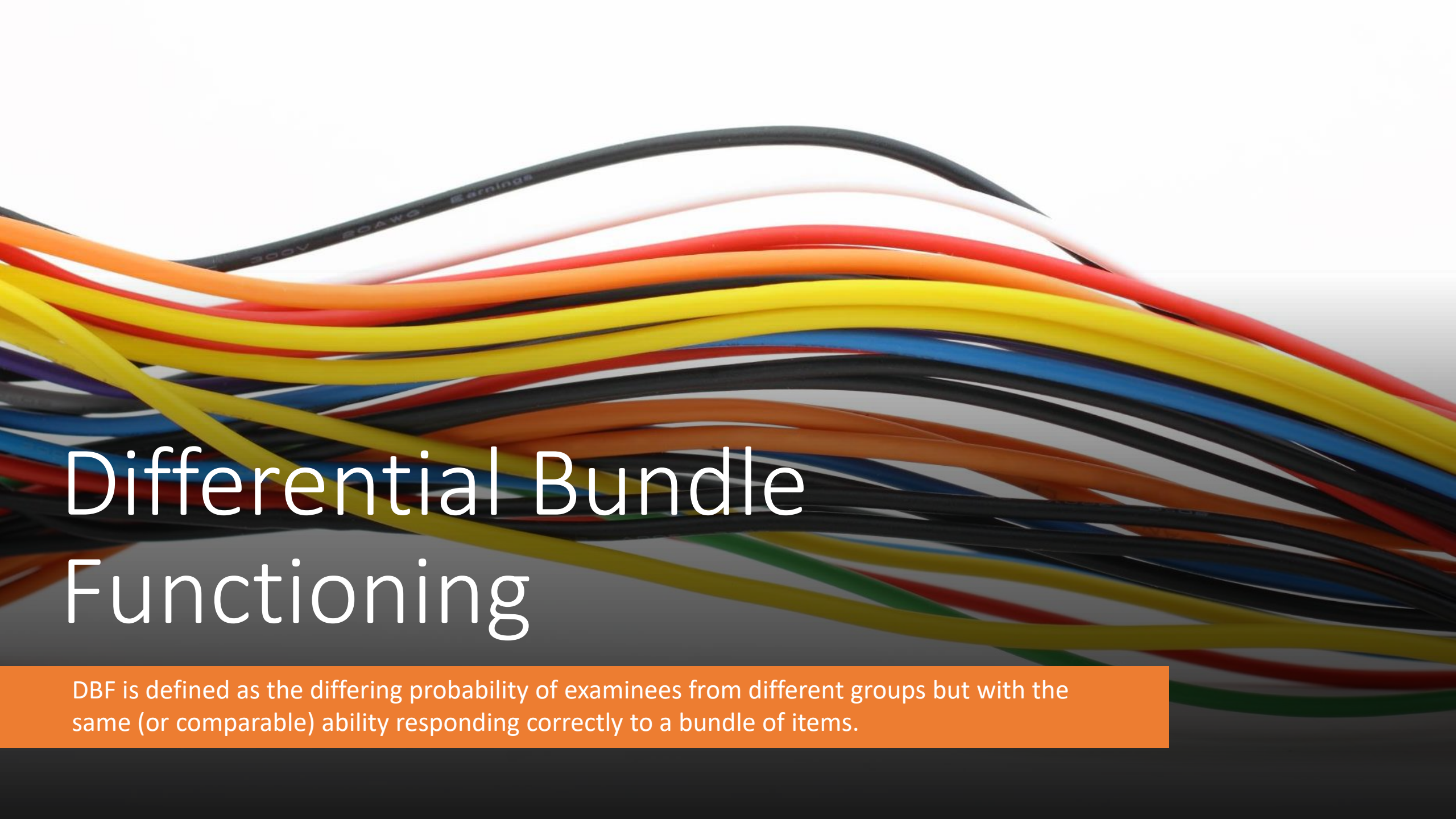
Differential item functioning

- DIF is “an unexpected difference among groups of examinees who are supposed to be comparable with respect to the attribute measured” (Dorans & Holland, 1993, p. 37).
- Large and statistically significant DIF raises concerns about multidimensionality, and it has to be taken as a violation of measurement invariance.
- Nonetheless, the differences in the test performance or in relation to specific items may represent ‘real’ difference in the mathematical construct being measured and may not indicate bias.
 - In this case, DIF is NOT disruptive for measurement. Yet, it can be used to quantify, understand and as a tool to fight inequalities.

DIF within the framework of the Rasch analysis

- In the Item Response Theory (IRT) methods, the absence of DIF in an item is defined as occurring when ICCs across different groups are identical (Hambleton & Rogers, 1989).



A bundle of many colorful, wavy lines in shades of yellow, orange, red, blue, and black, flowing from the left towards the right. The lines are of varying thickness and overlap each other, creating a sense of depth and movement. The background is a light gray gradient.

Differential Bundle Functioning

DBF is defined as the differing probability of examinees from different groups but with the same (or comparable) ability responding correctly to a bundle of items.



Literature review

- DB: Education Resources Information Center (ERIC)
- Keywords: 'Differential' AND 'Bundle' AND 'functioning'
- No time constraints
- PRISMA model
 - 28 publications (+ snowballing search) -> 34 publications (most of them aimed at proposing new methods to detect DBF).

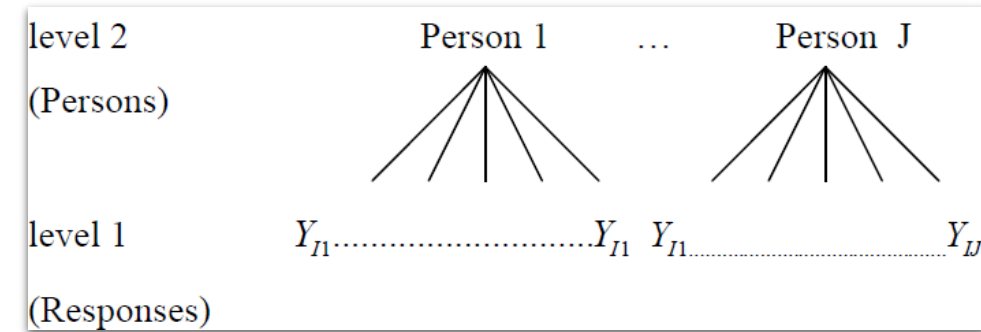
Studies analysing DF at **item-bundle** level using methods (other than SIBTEST)



Study	Methodology
Wainer et al. (1991)	Differential testlet functioning parallels to IRT likelihood procedures
Oshima et al. (1998)	Differential functioning of items and tests framework (DFIT)
Xie and Wilson (2008)	Differential facet functioning, an extension of Linear Logistic Test model Multidimensional differential facet functioning
Liu et al. (2008)	Multidimensional Rasch analysis
Swanson et al (2002)	Hierarchical logistic regression model



The basic model



$$\text{Logit} [\text{Prob}(Y_{ij}=1)] = b_{0j} + b_{1j} * \text{proficiency}_i + b_{2j} * \text{group}_i$$

- **proficiency** is an index of proficiency on a common scale for all examinees (rescaled to $N[0;1]$);
- **group_i** is a dichotomous var (group=0 for the reference group, and group =1 for the focal group);
- **b_{0j}** reflects (the log odds of) item difficulty in the reference group;
- **b_{1j}** reflects item discrimination (equal in reference and focal groups);
- **b_{2j}** reflects the deviation of item difficulty in the focal group from the reference group.



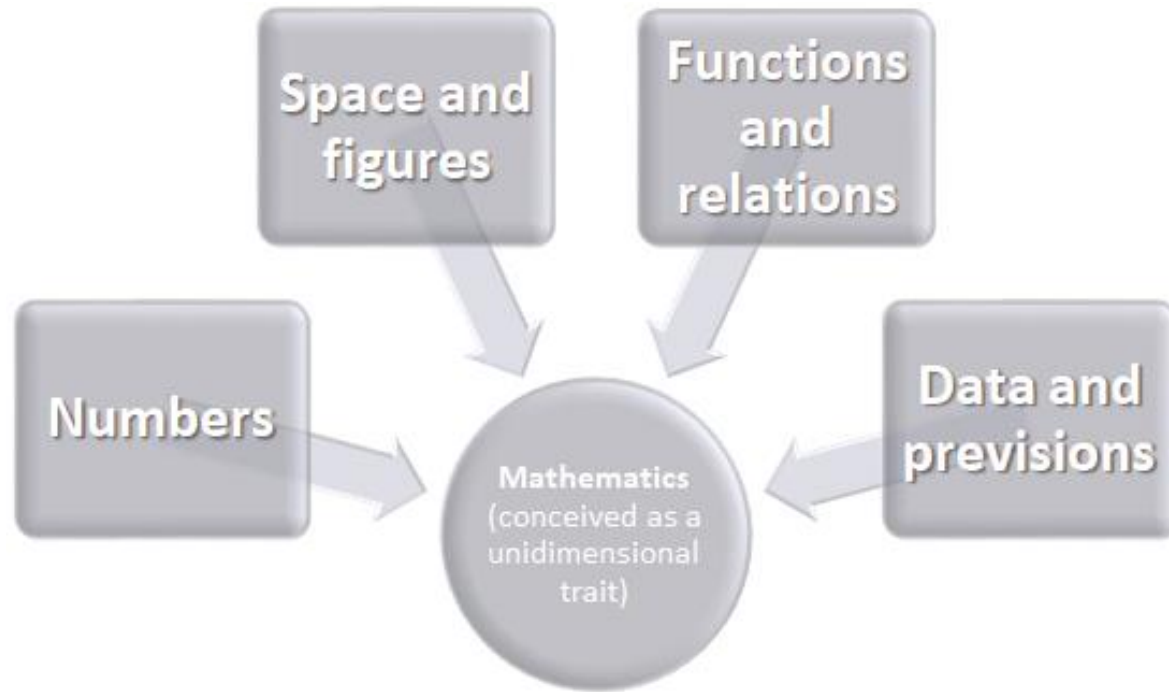
Data



Sample

- Project name: '**SIM12-16**'
 - Pilot study
 - Primary school, grade 5
 - 475 students (on average 10-years old students)

ZONE	NEIGHBOURHOOD	NUMBER OF SCHOOLS
A	Torbellamonaca	3
B	San Giovanni/Via Cavour	4
C	EUR/Trastevere	4
D	Parioli	2
E	Talenti	4
F	Cinecittà/Centocelle	3
		20



INVALSI Mathematics achievement tests



Results

STEP 1. DIF analysis

STEP 2. Bundles of items (and
interpretative hypotheses)

STEP 3. Hierarchical
modelling

Step 1

DIF analysis

DIF size (magnitude)	Classification
< 0.43	Negligible
0.44 < DIF < 0.64	Moderate
> 0.64	Large

ETS criteria (Zwick, 1999; 2002)

	ITEM	DIF SIZE			Sign
		Boy	Girls	Difference in absolute value	
1	A16	0,63	-1,29	1,92	**
2	A14_b	-0,56	0,44	1,00	*
3	A14_a	0,38	-0,54	0,92	*
4	A8	-0,32	0,58	0,90	*
5	A18	-0,33	0,57	0,90	*
6	A17	0,35	-0,46	0,81	
7	A21	-0,32	0,46	0,75	
8	A10	0,30	-0,44	0,64	
9	A25	-0,25	0,48	0,63	
10	A32	-0,31	0,38	0,69	
11	A14_c	0,27	-0,40	0,67	
12	A7	-0,27	0,29	0,56	
13	A20	-0,25	0,30	0,55	
14	A26	0,18	-0,33	0,51	
15	A22	0,18	-0,29	0,47	
16	A19_a	-0,20	0,26	0,46	
17	A11	0,17	-0,27	0,44	
18	A29	0,18	-0,26	0,34	
19	A33	0,15	-0,25	0,30	
20	A19_b	-0,18	0,21	0,39	
21	A31	0,12	-0,19	0,31	
22	A15	-0,13	0,17	0,30	
23	A5	0,12	-0,18	0,30	
24	A2	0,11	-0,15	0,26	
25	A4	-0,11	0,15	0,26	
26	A13	0,10	-0,15	0,25	
27	A30	0,10	-0,15	0,25	
28	A27	-0,10	0,14	0,24	
29	A12	0,09	-0,13	0,22	
30	A9	-0,09	0,11	0,20	
31	A3	0,07	-0,10	0,17	
32	A1	-0,07	0,06	0,13	
33	A28	0,05	-0,06	0,11	
34	A6	0,03	-0,05	0,08	
35	A24	0,03	-0,05	0,08	
36	A23	0,00	0,02	0,02	

Overall gender differences
(TEST SCORES) = less than
a quarter of standard
deviation (but statistically
significant)

A hand is shown in the top right corner, placing a single sheet of white paper on top of a tall stack of papers. The stack is part of a series of five stacks of increasing height from left to right. The background is a plain, light-colored wall and floor.

STEP 2: Bundles of items

Step 2: Bundling criteria (examples)

1. Item phrasing (words count)
 - HP: Longer items are more difficult to boys than to girls.
2. 'Scholastic' items (didactical contract)
 - HP: items more consistent with didactical praxes and textbooks are more difficult to girls than to boys.





Step 3

Differential bundle functioning:
hierarchical analysis to test
interpretative hypotheses

Results

Proficiency estimates are Rasch-based but rescaled to $N(0, 1)$;

The female dummy code was grand-mean centered. Taken together, these produce intercepts that are equal to the log odds of a correct response for examinees with a proficiency of zero.

All fixed effects are significantly different from 0 ($p < 0.005$), and all variance components are significantly greater than 0 ($p < 0.001$).



Mean increase in the log odds of a correct response associated with a 1-SD increase in proficiency

Mean increase in the log odds of a correct response for female examinees

Between-item variability in intercepts

Between-item variability in proficiency coefficients

Between-item variability in DIF coefficients

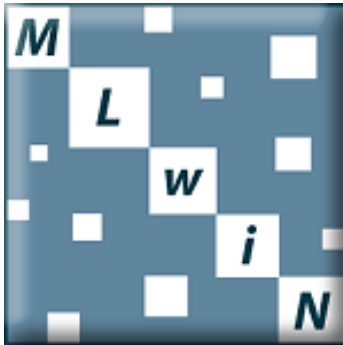
FIXED EFFECT	Regression coefficient	SE
Intercept	1.6547	0.0172
Proficiency	0.4826	0.0083
Girl (Ref: Boy)	0.0341	0.0084
RANDOM EFFECT	SD	Variance component
Intercept	1.5941	1.1423
Proficiency	0.1842	0.0297
Girl (Ref: Boy)	0.2316	0.0599

Results

(words count)

Interpretative hypothesis number 1 (the longer the text, the easier the mathematics item to girls) (e.g., Ajello et al., 2018; Cascella, 2021)

Change in the log odds of a correct response for female examinees for each increase of one word in the item

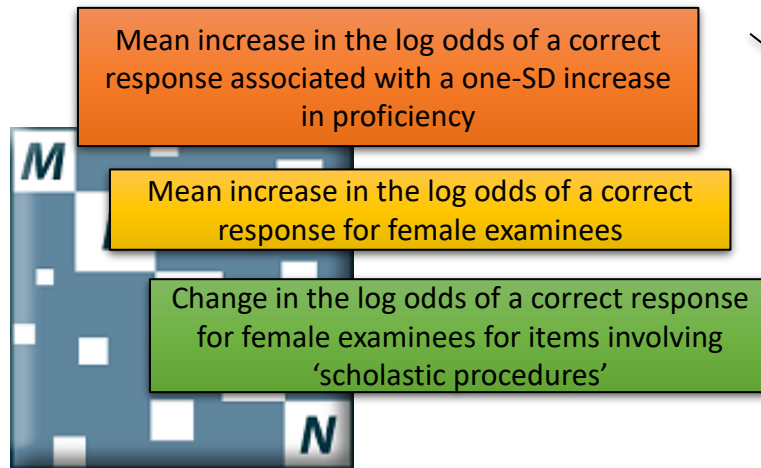


FIXED EFFECT	Regression coefficient	S.E.
Intercept	1.5547	0.0148
Proficiency	0.3626	0.0073
Girl (Ref: Boy)	0.0341	0.0084
Words count	0.0924	0.0141
RANDOM EFFECT	SD	Variance component
Intercept	1.0742	1.1245
Proficiency	0.1642	0.0295
Girl (Ref: Boy)	0.1914	0.0484

Results

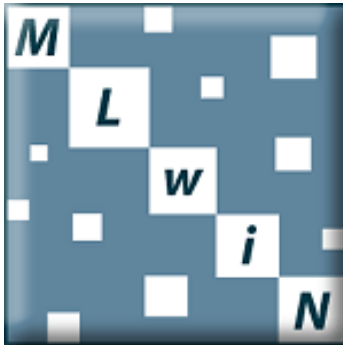
('scholastic' items)

Interpretative hypothesis number 2 (scholastic items are easier to girls than to boys)

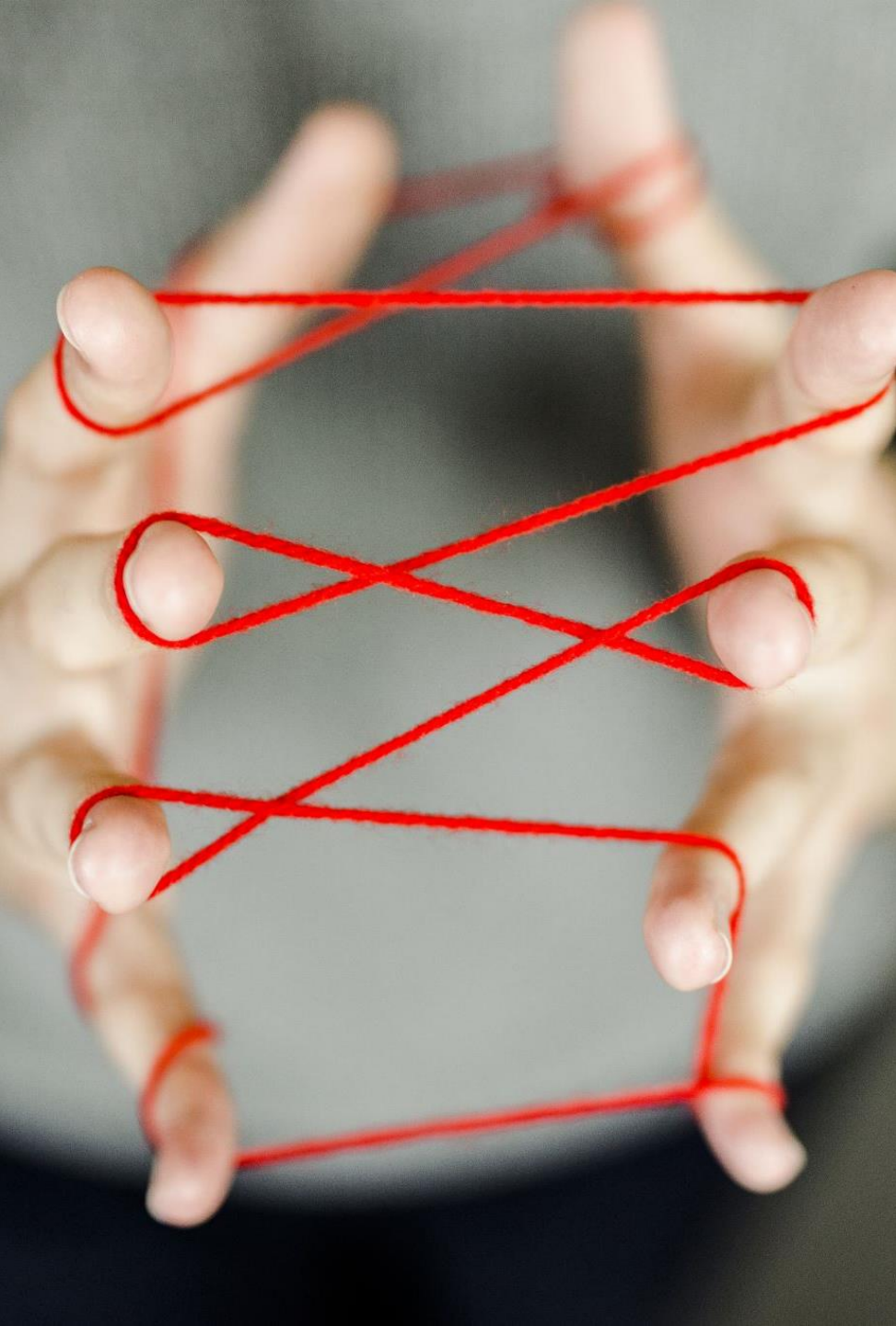


FIXED EFFECT	Regression coefficient	S.E.
Intercept	1.5678	0.0092
Proficiency	0.3567	0.0072
Girl (Ref: Boy)	0.0341	0.0084
'scholastic' item	0.1862	0.0083
RANDOM EFFECT	SD	Variance component
Intercept	1.0641	1.1321
Proficiency	0.1654	0.0308
Girl (Ref: Boy)	0.2064	0.0435

Results (overall)



FIXED EFFECT	Regression coefficient	S.E.
Intercept	1.3286	0.0197
Proficiency	0.3489	0.0051
Girl (Ref: Boy)	0.0123	0.0059
Word count	0.0511	0.0099
Scolastic items	0.1714	0.0240
RANDOM EFFECT	SD	Variance component
Intercept	1.0567	1.1324
Proficiency	0.1687	0.0345
Girl (Ref: Boy)	0.2002	0.4563



Conclusions

- Test scores show (often small and sometimes statistically not significant) gender differences, but do not provide any interpretative information.
- Investigation at item level can be used to identify items' characteristics associated with gender differences.
- Hierarchical logistic regression makes it possible:
 - to identify consistent sources of DIF across test items;
 - to quantify the proportion of explained variation in DIF coefficients;
 - to compare the predictive accuracy of alternate explanations for DIF.



Thanks for listening

Clelia Cascella
clelia.cascella@invalsi.it