

Differential item functioning across test versions

MACIEJ KONIEWSKI, PRZEMYSŁAW MAJKUT, PAULINA SKÓRSKA

Educational Research Institute



CONTEXTUAL INFORMATION

In Poland national exams are administered to students finishing primary (13 y.o.), lower secondary (16 y.o.), and higher secondary schools (19-21 y.o.).

The first high stake exam is being administered after lower secondary school. In this pen and paper exam the test is delivered in two booklet versions to reduce the risk of cheating.

Both versions share the same items, but the sequence of options for the multiple-choice items (MCQ) differs across versions.

OBJECTIVE

This study aims to assess differential item functioning (DIF) across the booklet versions, focusing on items grouped in testlets.

HYPOTHESIS

Differential functioning of items grouped in testlets occurs due to different correct answer patterns across booklet versions.

SAMPLE

DIF analyses across different booklet versions (A and B) were performed on 2013 data from 'The history and knowledge about society' test. The test was administered in lower secondary schools in three voivodeships: Lubelskie, Malopolskie and Podkarpackie (n=81 545).



The test consisted of 24 binary coded items out of which 15 were single items and 9 were grouped into 3 testlets.

METHODS

Mantel-Haenshel (M-H) test, logistic regression and standardization were used to detect and visualize DIF.

Logistic regression was used to test whether differential functioning of items grouped in testlets occurs due to different correct answer patterns across booklet versions.

'ANTI-PATTERN'

The testlets were presented to the examinees as shown below:

Text for task 11. ---steam---steam---steam---steam---
steam---steam---steam---steam---steam---steam---steam---

Instruction. ---instruction---instruction---instruction---
instruction---instruction---instruction---instruction---inst

---some text---some text--- **z11.1** ---some text---some
text--- **z11.2** ---some text---some text--- **z11.3** ---

z11.1 A. ---answer **B.** ---answer **C.** ---answer
z11.2 A. ---answer **B.** ---answer **C.** ---answer
z11.3 A. ---answer **B.** ---answer **C.** ---answer

In testlet z11 in booklet version B the correct answer pattern (in red) was marked with the same symbol (C, C, C).

Testlet z11 booklet version A			Testlet z11 booklet version B		
z11.1	A. B. C.		z11.1	A. B. C.	
z11.2	A. B. C.		z11.2	A. B. C.	
z11.3	A. B. C.		z11.3	A. B. C.	

Similar situation occurred in testlet z18 in booklet version A. The correct answer pattern was marked with A, A, A.

Testlet z18 booklet version A			Testlet z18 booklet version B		
z18.1	A. B. C.		z18.1	A. B. C.	
z18.2	A. B. C.		z18.2	A. B. C.	
z18.3	A. B. C.		z18.3	A. B. C.	

RESULTS OF DIF ANALYSIS

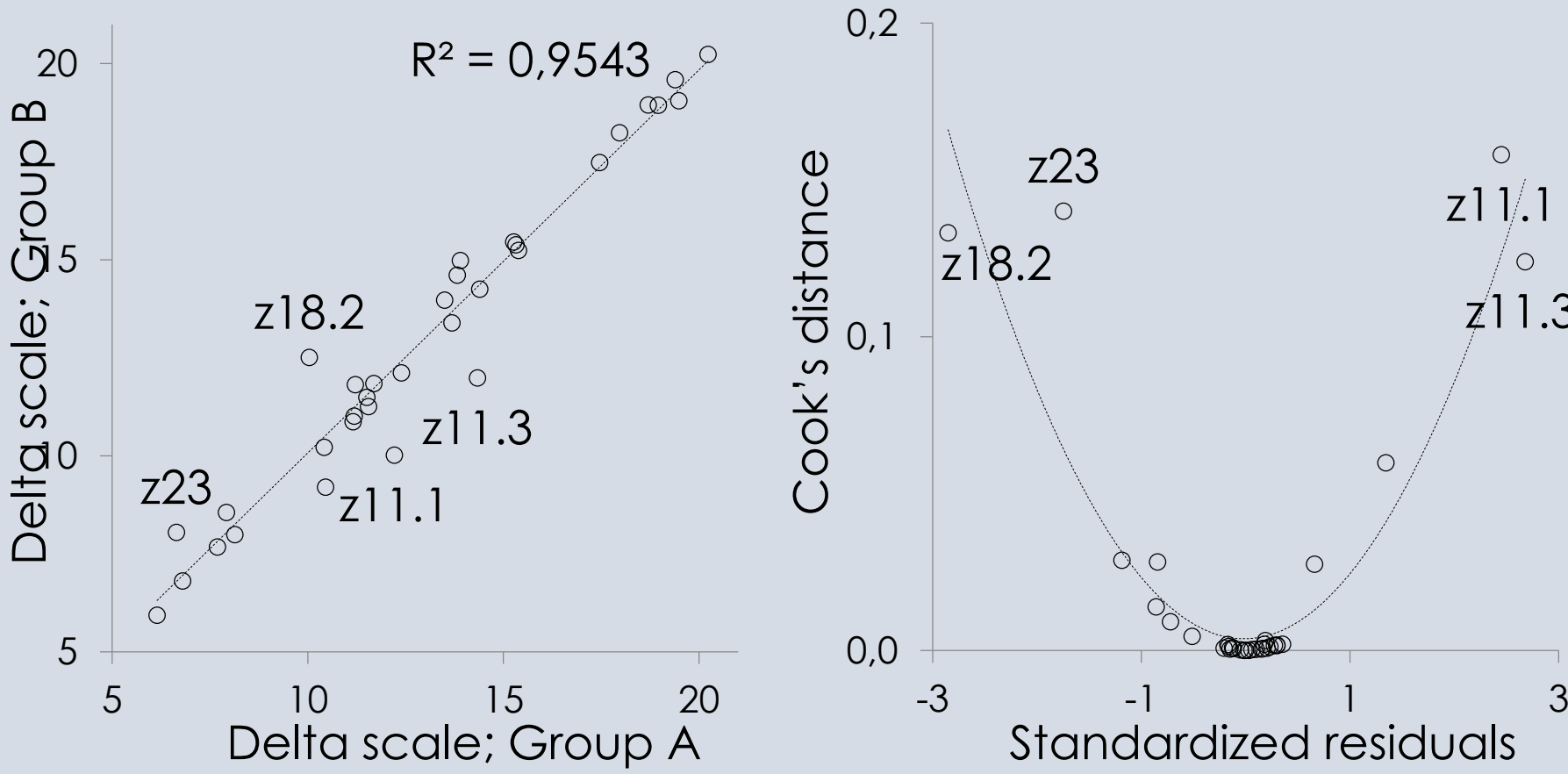
Four out of 24 items were identified as biased. Three biased items were grouped in testlets. Table below shows proportion of correct answers in focal and reference group.

	Booklet version A				Booklet version B				Diff.
	A	B	C	D	A	B	C	D	
z11.1	0,22	0,24	0,55	-	0,21	0,36	0,43	-	0,12
z11.2	0,02	0,91	0,07	-	0,02	0,08	0,91	-	0,00
z11.3	0,25	0,09	0,66	-	0,30	0,16	0,54	-	0,12
z18.1	0,98	0,02	0,00	-	0,98	0,02	0,00	-	0,00
z18.2	0,43	0,42	0,15	-	0,31	0,57	0,12	-	-0,13
z18.3	0,64	0,3	0,06	-	0,24	0,06	0,70	-	-0,06
z23	0,26	0,43	0,06	0,25	0,33	0,07	0,38	0,21	-0,08

Correct answer

The largest differences were observed for items grouped in testlets, where the correct answer pattern was marked with the same symbol in one booklet version (e.g., A, A, A) and in the alternative booklet version the symbols varied (e.g., A, B, C). This observation is in line with 'anti-pattern hypothesis'.

Based on transformed item-difficulty (TID) plot, Cook's distance and ZRESID plots the same items were identified as biased.



Further analysis focused on biased items grouped in testlets. According to Zwick and Ercikan (1989) classification effect sizes for these items are moderate.

Item	χ^2_{MH}	$\hat{\alpha}_{MH}$	$\ln(\hat{\alpha}_{MH})$	Δ_{MH}
z11.1	1390,44**	0,55	-0,59	1,40
z11.3	1529,21**	0,55	-0,60	1,40
z18.2	1622,67**	1,83	0,60	-1,42

** p<0,01

Uniform DIF was observed for item z11.1, while nonuniform DIF for items z11.3 and z18.2.

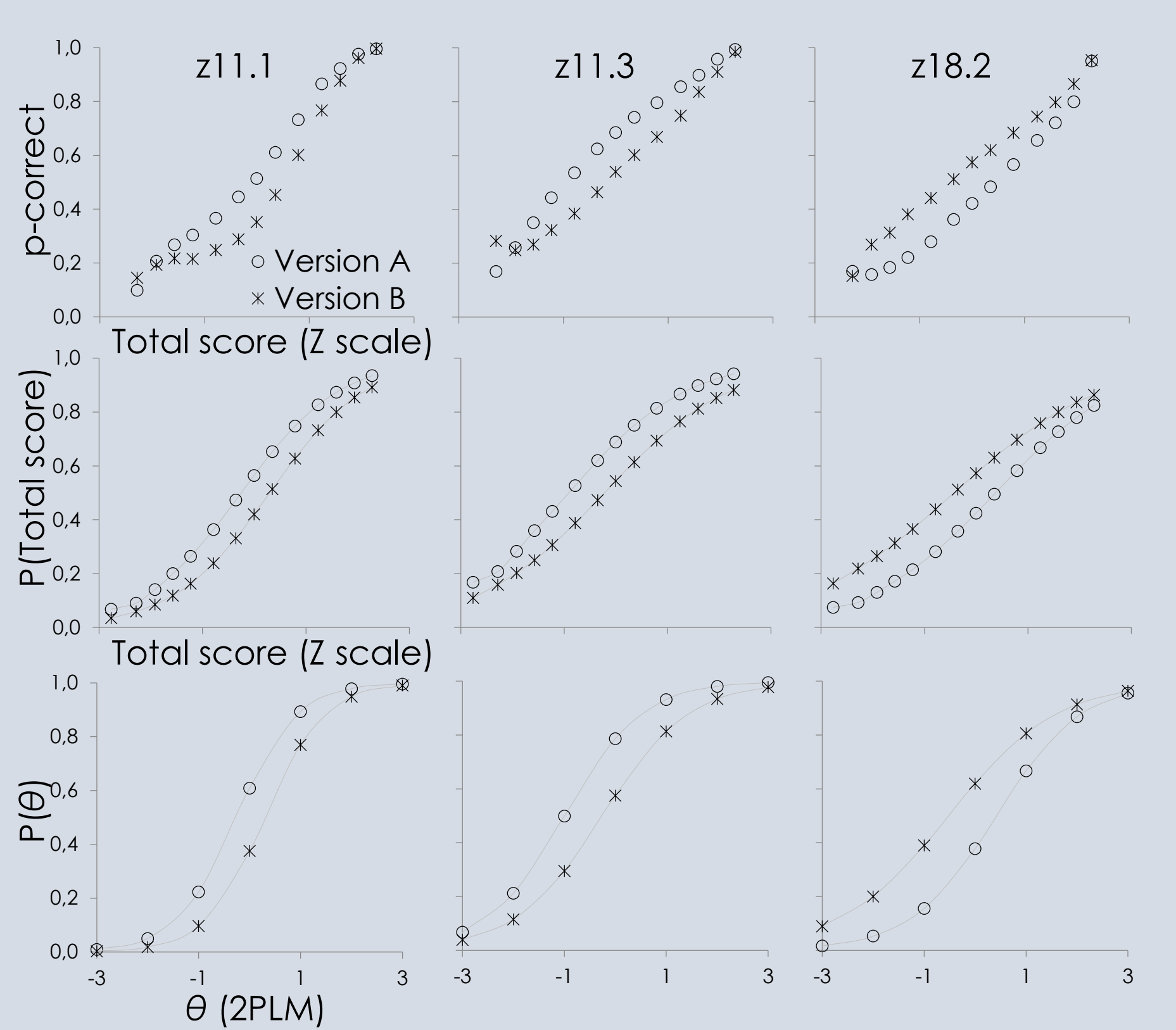
Item	2(LL M3 - LL M2)	2(LL M2 - LL M1)
z11.1	1,8	1392,6**
z11.3	12,4**	1547,8**
z18.2	54,8**	1614,8**

** p<0,01

$$\ln\left(\frac{P(u=1)}{P(u=0)}\right) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG$$

M1
M2
M3

X – total score
G – group of membership (focal or reference)
XG – interaction of total score and group membership.



'ANTI-PATTERN' HYPOTHESIS TESTING

Hypothesis of lower odds of correct answers on items in testlets where patterns were marked with the same symbol was tested. The following model was used:

$$\ln\left(\frac{P(u=1)}{P(u=0)}\right) = \beta_0 + \beta_1 X + \beta_2 Z$$

X – total score
Z – correct answers for two remaining items in testlet.

		β	SE	Exp(β)
z11.1	Test score (-z11.1)	0,172**	0,002	1,188
	Correct answers on both z11.2 & z11.3	-0,388**	0,024	0,678
	Intercept-3,408**		0,045	0,033
z11.3	Test score (-z11.3)	0,123**	0,002	1,131
	Correct answers on both z11.1 & z11.2	-0,315**	0,025	0,730
	Intercept-2,053**		0,041	0,128
z18.2	Test score (-z18.2)	0,136**	0,002	1,146
	Correct answers on both z18.1 & z18.3	-0,731**	0,024	0,481
	Intercept-2,436**		0,040	0,088

** p<0,01

CONCLUSIONS

Results indicated meaningful differences in items functioning across booklet versions. The largest differences occurred when the correct answer pattern was marked with the same symbol (e.g., A, A, A). Such a pattern has been named an 'anti-pattern', because test takers might consider such response pattern very unlikely and provide wrong answers in consequence.

The two booklet versions are theoretically parallel in terms of test items. However, differences in answer patterns might influence students' performance on the items.

Controlling for total test score, correct answers provided for two items in a testlet decrease the odds of correct answer on the remaining item by 27 up to 52%.

REFERENCES

- Angoff, W. H. (1972) *The development of statistical indices for detecting cheaters*. Princeton, NJ: ETS.
- Dorans, N. J., Kulick, E. (1984). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355-368.
- Holland, P.W. & Wainer, H. (ed.) (2009) *Differential Item Functioning*. New York: Routledge.
- 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.
- Swaminathan, H., Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- Wainer, H. & Braun H. (ed.) (1998) *Test validity*. Hillsdale, NJ: Erlbaum
- Zwick, R. & Ercikan, K. (1989) Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, 26, 55-66.

@ m.koniewski@ibe.edu.pl | maciej.koniewski@uj.edu.pl
p.majkut@ibe.edu.pl | p.majkut@uj.edu.pl
p.skorska@ibe.edu.pl | paulina.skorska@uj.edu.pl