Supplemental Material

CBE—Life Sciences Education

Martinková *et al.*

Martinková P, Drabinová A, Liaw Y-L, Sanders EA, McFarland JL, Price RM. Using DIF analysis to reveal potential equity gaps in conceptual assessments

Supplemental Materials, Table 1. Test statistics and p-values of DIF detection methods for HCI data set. All p-values are >0.05, no item is detected as DIF item. 3PL IRT model was fitted on dataset without problematic Item 17. Small power (most p-values close to 1) for the IRT-based method is in line with simulation studies (Kim and Oshima, 2013, Drabinová and Martinková, in review): IRT-based methods need at least 500 respondents in each group to function well.

Item	Mantel-Haenszel					Logistic regression				3 PL IRT Wald's test	
						2 PL		3 PL			
	χ^2_{MH}	p-value	α_{MH}	Δ_{MH}		LRT	p-value	F	p-value	χ^2	p-value
1	4.14	0.51	1.52	-0.99	A	6.54	0.21	3.47	0.13	8.93	0.11
2	0.02	0.97	0.95	0.13	A	0.12	0.94	0.04	0.96	0.49	0.88
3	0.08	0.91	1.13	-0.28	A	0.40	0.94	0.36	0.82	0.96	0.88
4	2.36	0.59	1.34	-0.70	A	3.03	0.55	1.71	0.42	2.96	0.62
5	0.54	0.91	0.86	0.34	A	1.36	0.85	1.90	0.42	1.83	0.82
6	1.43	0.72	0.78	0.58	A	2.69	0.58	1.71	0.42	0.36	0.88
7	0.52	0.91	0.87	0.33	A	0.38	0.94	0.19	0.92	0.38	0.88
8	0.81	0.91	0.81	0.49	A	1.72	0.77	0.72	0.73	1.53	0.82
9	2.09	0.59	0.76	0.65	A	3.33	0.55	1.24	0.53	2.99	0.62
10	1.31	0.72	0.79	0.57	A	2.43	0.59	1.66	0.42	0.37	0.88
11	0.01	0.99	0.99	0.02	A	5.00	0.33	4.40	0.10	2.56	0.66
12	2.46	0.59	1.37	-0.74	A	6.32	0.21	3.80	0.11	3.49	0.62
13	0.11	0.91	0.92	0.20	A	3.23	0.55	0.53	0.79	0.38	0.88
14	0.10	0.91	0.92	0.22	A	0.26	0.94	0.09	0.96	0.12	0.94
15	0.30	0.91	1.13	-0.28	A	0.43	0.94	0.68	0.73	4.18	0.59
16	0.29	0.91	1.13	-0.29	A	0.21	0.94	0.67	0.73	1.48	0.82
17	0.00	0.99	1.02	-0.03	A	0.55	0.94	1.41	0.49	-	-
18	0.09	0.91	0.89	0.27	A	0.27	0.94	0.39	0.82	0.94	0.88
19	3.80	0.51	1.62	-1.14	A	9.70	0.10	4.22	0.10	8.07	0.11
20	0.25	0.91	0.88	0.30	A	9.18	0.10	4.33	0.10	8.05	0.11

Supplemental Materials, Table 2. Steps used to generate simulated dataset.

1. Generate latent trait from normal distribution (with zero mean and SD=1) for 1,000 men and 12,000 women
2. Select item parameters for 20 items.
    a. The first item was manipulated to have uniform DIF with parameters $a = 1,\ b = 0,\ b_{DIF} = 1$; the second, to have non-uniform DIF with parameters $a = 0.56,\ b = 0,\ a_{DIF} = 1.23$ and $b_{DIF} = 0.5$. For both DIF items, the guessing parameter c was set to 0.2, meaning that students obtained the correct answer through guessing 20% of the time.
    b. To reflect realistic values of items' parameters, we used the estimated parameters' values of first 18 items of Graduate Management Admission Test (GMAT, Kingston et al., 1985, p. 47), see Table 2 below.
3. The probabilities of correct answers were calculated from two separate 3PL IRT models, one for men (reference group) and one for women (focal group).
4. The 0/1 responses of examinees were generated from Bernoulli distribution with calculated probabilities.
5. Total score was calculated for each student.
6. We paired each man with a woman who had the same total score. That woman was randomly picked from the remaining women with the same total score. The remaining 11,000 women were removed from the sample.

Martinková P, Drabinová A, Liaw Y-L, Sanders EA, McFarland JL, Price RM. Using DIF analysis to reveal potential equity gaps in conceptual assessments

Supplemental Materials, Table 3. True parameters applied in the simulated dataset based on GMAT. The first two items are manipulated to be DIF items: Item 1 is manipulated to have uniform DIF (only b_{DIF} parameter is non-zero, there is a shift in difficulty for the two groups), Item 2 is manipulated to have non-uniform DIF (a_{DIF} is non-zero, the item has different slope and also b_{DIF} is non-zero, the item has a different difficulty). Values in **bold** come from GMAT (Kingston et al., 1985, p. 47).

Item	a	b	c	a_{DIF}	b_{DIF}
1	1.00	0.00	0.20	0.00	1.00
2	0.56	0.00	0.20	1.23	0.50
3	**0.29**	**-2.95**	**0.07**	0.00	0.00
4	**0.41**	**-2.93**	**0.07**	0.00	0.00
5	**0.94**	**-1.21**	**0.33**	0.00	0.00
6	**0.88**	**-0.24**	**0.18**	0.00	0.00
7	**0.42**	**-1.15**	**0.07**	0.00	0.00
8	**0.74**	**0.60**	**0.36**	0.00	0.00
9	**0.35**	**-0.35**	**0.07**	0.00	0.00
10	**0.44**	**-0.30**	**0.07**	0.00	0.00
11	**0.55**	**-1.06**	**0.07**	0.00	0.00
12	**0.82**	**1.02**	**0.36**	0.00	0.00
13	**0.52**	**-1.96**	**0.07**	0.00	0.00
14	**1.02**	**1.28**	**0.22**	0.00	0.00
15	**0.65**	**0.49**	**0.16**	0.00	0.00
16	**0.82**	**0.61**	**0.07**	0.00	0.00
17	**1.04**	**2.11**	**0.37**	0.00	0.00
18	**0.95**	**0.81**	**0.09**	0.00	0.00
19	**1.01**	**0.81**	**0.19**	0.00	0.00
20	**0.98**	**1.67**	**0.28**	0.00	0.00

Martinková P, Drabinová A, Liaw Y-L, Sanders EA, McFarland JL, Price RM. Using DIF analysis to reveal potential equity gaps in conceptual assessments

Supplemental Materials, Table 4. Test statistics and p-values of DIF detection methods for simulated data set. The first two items are detected as DIF by all methods (corrected p-value < 0.01). On delta scale, Item 1 is classified within Category C (large difference), while Item 2 is classified into Category A (negligible difference), pointing to limited ability of MH test to detect non-uniform DIF.

Item	Mantel-Haenszel					Logistic regression				3 PL IRT Wald's test	
						2 PL		3 PL			
	χ^2_{MH}	p-value	α_{MH}	Δ_{MH}		LRT	p-value	F	p-value	χ^2	p-value
1	69.84	<0.01	2.29	-1.94	C	73.52	<0.01	41.78	<0.01	42.53	<0.01
2	14.36	<0.01	1.45	-0.87	A	27.56	<0.01	14.23	<0.01	16.30	<0.01
3	0.52	0.72	0.92	0.19	A	0.74	0.79	0.34	0.84	0.91	0.79
4	2.62	0.35	0.83	0.44	A	3.41	0.46	1.65	0.52	2.16	0.68
5	0.00	1.00	0.99	0.02	A	0.69	0.79	0.60	0.73	0.18	0.96
6	0.06	0.86	1.03	-0.07	A	0.11	0.95	0.08	0.93	0.00	1.00
7	6.56	0.07	0.77	0.60	A	7.21	0.18	4.15	0.11	5.30	0.35
8	0.23	0.84	0.95	0.12	A	4.97	0.33	1.43	0.53	3.68	0.64
9	0.06	0.86	0.97	0.06	A	0.43	0.85	0.23	0.89	1.65	0.68
10	0.89	0.62	0.91	0.22	A	1.36	0.78	0.67	0.73	1.37	0.70
11	0.26	0.84	1.06	-0.13	A	1.01	0.79	0.62	0.73	2.29	0.68
12	0.11	0.86	0.97	0.08	A	1.08	0.79	0.53	0.74	0.40	0.91
13	1.02	0.62	0.89	0.27	A	3.22	0.46	2.20	0.44	1.67	0.68
14	1.68	0.53	0.88	0.30	A	2.38	0.61	0.75	0.73	1.29	0.70
15	0.79	0.62	0.92	0.21	A	1.42	0.78	0.65	0.73	0.67	0.84
16	0.09	0.86	0.97	0.08	A	0.88	0.79	0.07	0.93	1.83	0.68
17	2.82	0.35	0.85	0.38	A	3.13	0.46	1.58	0.52	2.38	0.68
18	1.20	0.61	0.89	0.27	A	1.38	0.78	1.01	0.73	2.00	0.68
19	1.57	0.53	0.88	0.30	A	5.01	0.33	3.12	0.22	7.50	0.16
20	3.00	0.35	1.19	-0.40	A	3.76	0.46	1.74	0.52	2.25	0.68

Martinková P, Drabinová A, Liaw Y-L, Sanders EA, McFarland JL, Price RM. Using DIF analysis to reveal potential equity gaps in conceptual assessments

Supplemental Materials, Table 5. Estimated coefficients of characteristic curves by model NLR for HCI data set. a is discrimination parameter, b is difficulty parameter and c is the parameter that estimates guessing, an improvement that is standard in IRT, and advocated for in LR by Drabinová and Martinková (2016). Estimations are complemented by standard errors (s.e.). There is no significant DIF in any of the items, thus the final LR model has only 3 parameters for each item (see also Table 6 bellow, where DIF in first two items corresponds to non-zero fourth or fifth parameter).

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)
1	0.98	(0.25)	-1.03	(0.69)	0.00	(0.31)
2	0.94	(0.30)	-0.98	(0.88)	0.21	(0.30)
3	1.37	(0.30)	-1.67	(0.55)	0.00	(0.36)
4	1.63	(0.45)	1.24	(0.13)	0.25	(0.04)
5	1.14	(0.26)	0.56	(0.22)	0.10	(0.09)
6	1.24	(0.23)	0.65	(0.16)	0.01	(0.07)
7	0.48	(0.33)	-0.41	(3.00)	0.00	(0.69)
8	1.24	(0.27)	-0.65	(0.36)	0.15	(0.16)
9	1.46	(0.38)	1.02	(0.15)	0.23	(0.05)
10	2.36	(0.57)	0.38	(0.12)	0.42	(0.04)
11	1.06	(0.27)	-1.39	(0.76)	0.04	(0.37)
12	1.44	(0.30)	0.23	(0.19)	0.22	(0.08)
13	2.67	(0.53)	0.27	(0.09)	0.32	(0.04)
14	1.65	(0.35)	-0.54	(0.23)	0.31	(0.10)
15	1.76	(0.35)	0.71	(0.11)	0.19	(0.05)
16	1.44	(0.26)	-0.20	(0.20)	0.11	(0.09)
17	2.80	(0.88)	1.67	(0.12)	0.24	(0.02)
18	1.83	(0.31)	-1.20	(0.21)	0.00	(0.15)
19	1.85	(0.37)	-0.84	(0.20)	0.23	(0.11)
20	1.15	(0.25)	-1.05	(0.49)	0.00	(0.25)

Martinková P, Drabinová A, Liaw Y-L, Sanders EA, McFarland JL, Price RM. Using DIF analysis to reveal potential equity gaps in conceptual assessments

Supplemental Materials, Table 6. Estimated coefficients of characteristic curves by 3-paremeter logistic regression model for simulated data set. $\boldsymbol{a}$ is discrimination parameter, $\boldsymbol{b}$ is difficulty parameter and $\boldsymbol{c}$ is the parameter that estimates guessing, an improvement that is standard in IRT, and advocated for in LR by Drabinová and Martinková (2016). Estimations are complemented by standard errors (s.e.). $\boldsymbol{a_{DIF}}$ and $\boldsymbol{b_{DIF}}$ are differences in discrimination and difficulty parameters between reference and focal group. The first two items are detected as DIF, item 1 as uniform DIF and item 2 as non-uniform DIF.

Item	a	s.e.(a)	b	s.e.(b)	c	s.e.(c)	a_{DIF}	s.e.(a_{DIF})	b_{DIF}	s.e.(b_{DIF})
1	1.12	(0.15)	-0.30	(0.17)	0.11	(0.07)	-0.06	(0.17)	0.95	(0.11)
2	0.72	(0.10)	-0.48	(0.30)	0.08	(0.08)	0.53	(0.18)	0.50	(0.16)
3	0.67	(0.16)	-1.50	(1.21)	0.00	(0.41)	-	-	-	-
4	0.63	(0.18)	-2.19	(1.99)	0.00	(0.69)	-	-	-	-
5	1.07	(0.20)	-1.07	(0.45)	0.33	(0.15)	-	-	-	-
6	0.91	(0.15)	-0.60	(0.39)	0.06	(0.15)	-	-	-	-
7	0.68	(0.16)	-1.00	(0.90)	0.00	(0.29)	-	-	-	-
8	0.82	(0.17)	-0.15	(0.45)	0.19	(0.13)	-	-	-	-
9	0.64	(0.16)	-0.38	(0.82)	0.05	(0.23)	-	-	-	-
10	0.66	(0.16)	-0.26	(0.73)	0.08	(0.20)	-	-	-	-
11	0.86	(0.15)	-0.95	(0.52)	0.01	(0.20)	-	-	-	-
12	0.90	(0.17)	0.38	(0.28)	0.21	(0.09)	-	-	-	-
13	0.82	(0.19)	-0.77	(0.70)	0.29	(0.18)	-	-	-	-
14	0.88	(0.15)	0.54	(0.22)	0.07	(0.08)	-	-	-	-
15	0.60	(0.16)	-0.16	(0.83)	0.00	(0.23)	-	-	-	-
16	1.08	(0.15)	0.47	(0.14)	0.08	(0.06)	-	-	-	-
17	0.68	(0.17)	0.76	(0.40)	0.11	(0.12)	-	-	-	-
18	1.01	(0.14)	0.53	(0.15)	0.03	(0.06)	-	-	-	-
19	0.87	(0.14)	0.19	(0.26)	0.03	(0.10)	-	-	-	-
20	1.26	(0.19)	1.07	(0.10)	0.20	(0.04)	-	-	-	

Martinková P, Drabinová A, Liaw Y-L, Sanders EA, McFarland JL, Price RM. Using DIF analysis to reveal potential equity gaps in conceptual assessments

Supplemental Materials, Selected R code

```
# installation/loading of R packages

# install.library("difR")

library("difR")

# install.library("difNLR")

library("difNLR")


# data loading

data(GMAT)

data <- GMAT [, 1:20]

group <- GMAT [, "group"]


# DIF detection


# 1. Mantel-Haenszel X2 test

# calculation of MH statistics and MH estimates of odds ratio with difR package

# using Benjamini-Hochberg multiple comparison correction

(fitMH <- difMH(data, group, focal.name = 1, p.adjust.method = "BH"))

# printing which items are detected as DIF
```

```
fitMH$DIFitems
```

 # 2. Logistic Regression

```
# calculation of LR test statistic with difR package
# using Benjamini-Hochberg multiple comparison correction
(fitLR <- difLogistic(data, group, focal.name = 1, p.adjust.method = "BH"))
# printing which items are detected as DIF
fitLR$DIFitems
```

 # 3. Non-Linear Regression

```
# calculation of F test statistic with difNLR package
# using Benjamini-Hochberg multiple comparison correction (default option)
(fitNLR <- difNLR(data, group, focal.name = 1, model = "3PLcg", test = "F", p.adjust.method = "BH"))
# print characteristic curves for item 1
plot(fitNLR, item = 1)
```

 # 4. 3PL IRT – Wald's (Lord's) statistic

```
# estimating 3 PL IRT model for both groups with difR package
```

Martinková P, Drabinová A, Liaw Y-L, Sanders EA, McFarland JL, Price RM. Using DIF analysis to reveal potential equity gaps in conceptual assessments

```
# for HCI data set use data <- data[, -17]
fitIRTguess <- itemParEst(data, model = "3PL")
# extracting common guessing parameter for both groups
guess <- fitIRTguess [, 3]

# calculation of Lord's statistic with difR package
# using Benjamini-Hochberg multiple comparison correction
(fitLord <- difLord (data, group, focal.name = 1,
                     model = "3PL", c = guess, p.adjust.method = "BH"))
# printing which items are detected as DIF
fitLord$DIFitems



# Run interactive shiny application with GMAT data
# install.packages("ShinyItemAnalysis")
# application needs cleared workspace
rm(list = ls()) # clears the workspace
library("ShinyItemAnalysis")
startShinyItemAnalysis()
```