Methods for Improving the Goodness-of-fit By Considering Responses and Response Time

Heru Widiatmo and Lisa A. Gawlick

Background

For any chosen IRT model and any given test data, one or more items of the test might be found to be misfit.

The misfit items, however, might not always be due to poorly-performing items, but it might be due to the irregularities of the examinees' responses.

By excluding those examinees, therefore, the goodness-of-fit of the items could be improved.

In addition, taking care of the data irregularities could improve measurement precision of the tests (Wise & Kong, 2005; Meijer & Sotaridona, 2006; van der Linden, 2006; Marianti, Fox, Avetisyan, Veldkamp, 2014).

Background

Two methods considered and compared:

1. Person-Fit Statistics (*PFS*) (Levin & Rubin,1979; Drasgow, Levin, & Williams,1985; Meijer, Niessen, & Tendeiro, 2016)

   Based on an IRT model and responses (1/0 data)

2. Effective Response Time (*ERT*) (Meijer & Sotaridona, 2006)

   Based on an IRT model, responses, and response times (RTs)

Purpose : To investigate whether *PFS* and/or *ERT* methods can screen irregular examinees such that “clean” data can be obtained and then the goodness-of-fit can be improved.

A total of 24 (2x4x3) conditions were considered and evaluated against each other and to the benchmark:

- Two IRT models (*2-PL* and *3-PL*)
- Four methods: *ERT, PFS, ERT_PFS,* and *PFS_ERT*
- Three significance levels (α): 0.01, 0.05, and 0.10

Notation

PFS10 ≡ *PFS* method and α = 10%

ERT_PFS01 ≡ The methods were used together and α = 1%

- Data (2,734 examinees) came from a computerized test in an operational testing program for high school students
- The test consists of 60 MC items with five options

Table 1

Raw and Response Time Statistics

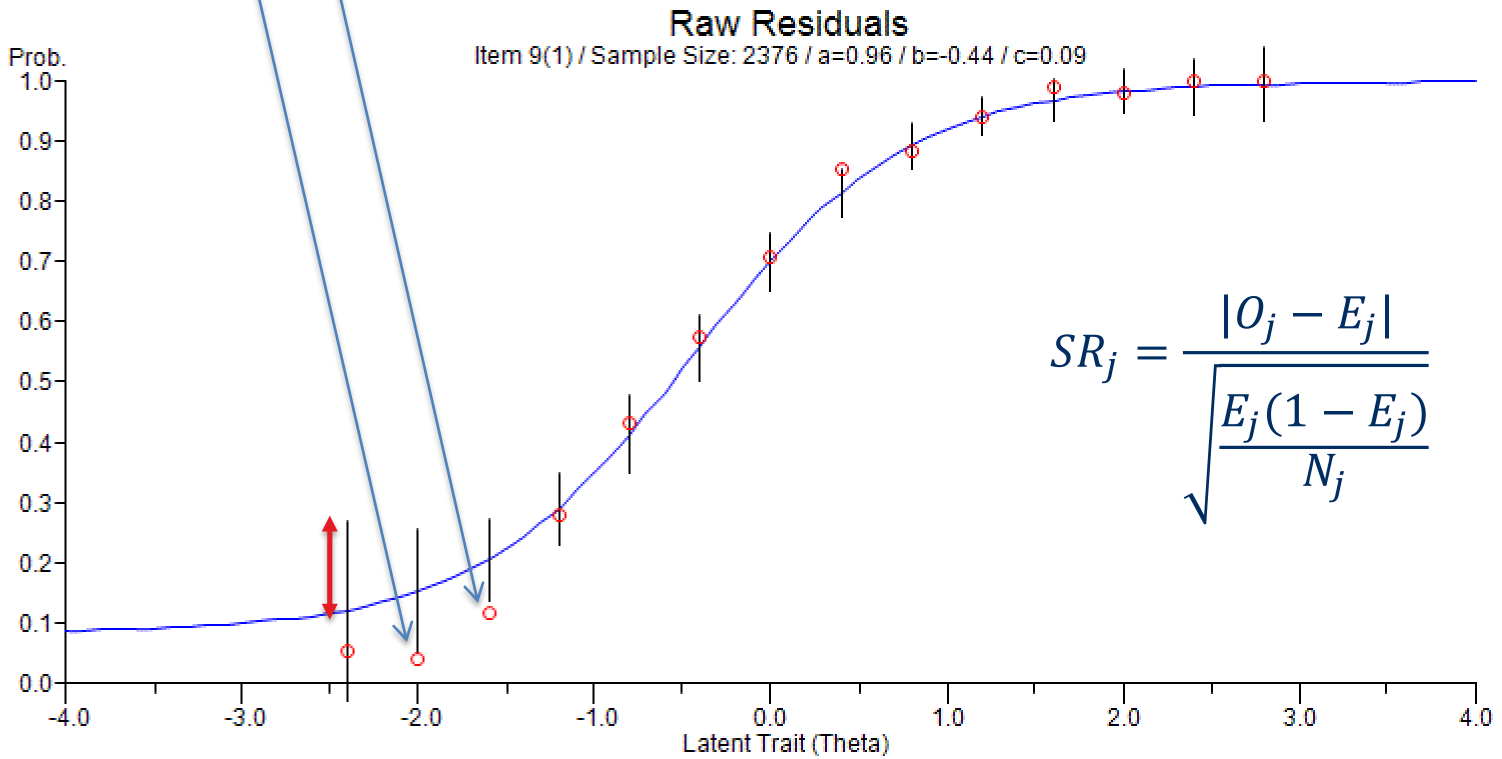
	Raw Score	RT (in second)
Average	27.37	3062
SD	10.15	460
Min	3	1505
Max	59	3564
N	2734	2734

46%

51 sec/item

Criteria Measured

The number of score misfits, absolute standardized residual, chi-square goodness-of-fit with α = 0.05, and estimated abilities



$$SR_j = \frac{|O_j - E_j|}{\sqrt{\frac{E_j(1 - E_j)}{N_j}}}$$

Results

Table 2
Number of Item Misfit

2-PL

		α=0.01				α=0.05				α=0.10			
Misfit	Bench.	ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT
0	13	14	13	14	14	17	14	17	16	18	14	17	17
1	19	27	19	27	27	23	18	23	25	23	20	22	22
2	13	8	13	8	8	9	13	10	8	8	11	12	11
3	8	5	8	5	5	3	6	2	3	5	7	3	2
4	2	3	2	3	3	4	3	4	4	4	3	3	6
5	2	0	2	0	0	1	4	1	1	1	3	3	1
6	2	2	2	2	2	3	1	3	3	1	1	0	1
7	1	1	1	1	1	0	1	0	0	0	1	0	0
N	2734	2372	2734	2372	2372	2218	2729	2216	2211	2118	2728	2112	2112

3-PL

		α=0.01				α=0.05				α=0.10			
Misfit	Bench.	ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT
0	26	27	27	27	29	27	27	27	32	31	27	27	27
1	20	18	18	21	21	23	20	21	20	22	16	21	23
2	8	11	9	7	7	4	8	7	5	4	10	7	6
3	3	2	4	4	1	4	4	5	2	2	4	5	3
4	2	2	1	0	2	2	1	0	1	1	3	0	1
5	1	0	1	1	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0
N	2734	2376	2728	2373	2372	2228	2714	2220	2218	2125	2691	2102	2088

Results

Table 3

Descriptive Statistics for Standardized Residual

2PL

Method	Bench.	α=0.01				α=0.05				α=0.10			
		ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT
Average	0.929	0.898	0.929	0.898	0.898	0.905	0.920	0.906	0.909	0.876	0.918	0.876	0.880
SD	0.269	0.237	0.269	0.237	0.237	0.240	0.256	0.241	0.240	0.234	0.249	0.232	0.235
Min	0.546	0.467	0.546	0.467	0.467	0.518	0.536	0.531	0.519	0.451	0.556	0.467	0.432
Max	1.749	1.597	1.749	1.597	1.597	1.672	1.767	1.682	1.680	1.650	1.642	1.660	1.664
N	2734	2372	2734	2372	2372	2218	2729	2216	2211	2118	2728	2112	2112

3PL

Method	Bench.	α=0.01				α=0.05				α=0.10			
		ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT
Average	0.800	0.788	0.793	0.780	0.791	0.775	0.789	0.780	0.770	0.762	0.790	0.778	0.761
SD	0.197	0.187	0.208	0.182	0.194	0.184	0.185	0.182	0.170	0.167	0.177	0.182	0.163
Min	0.485	0.467	0.387	0.387	0.445	0.451	0.449	0.387	0.418	0.461	0.480	0.445	0.453
Max	1.374	1.355	1.582	1.298	1.348	1.398	1.358	1.298	1.331	1.325	1.338	1.387	1.356
N	2734	2376	2728	2373	2372	2228	2714	2220	2218	2125	2691	2102	2088

Results

Table 4
Misfit Items Based on Chi-square Goodness of Fit
2PL

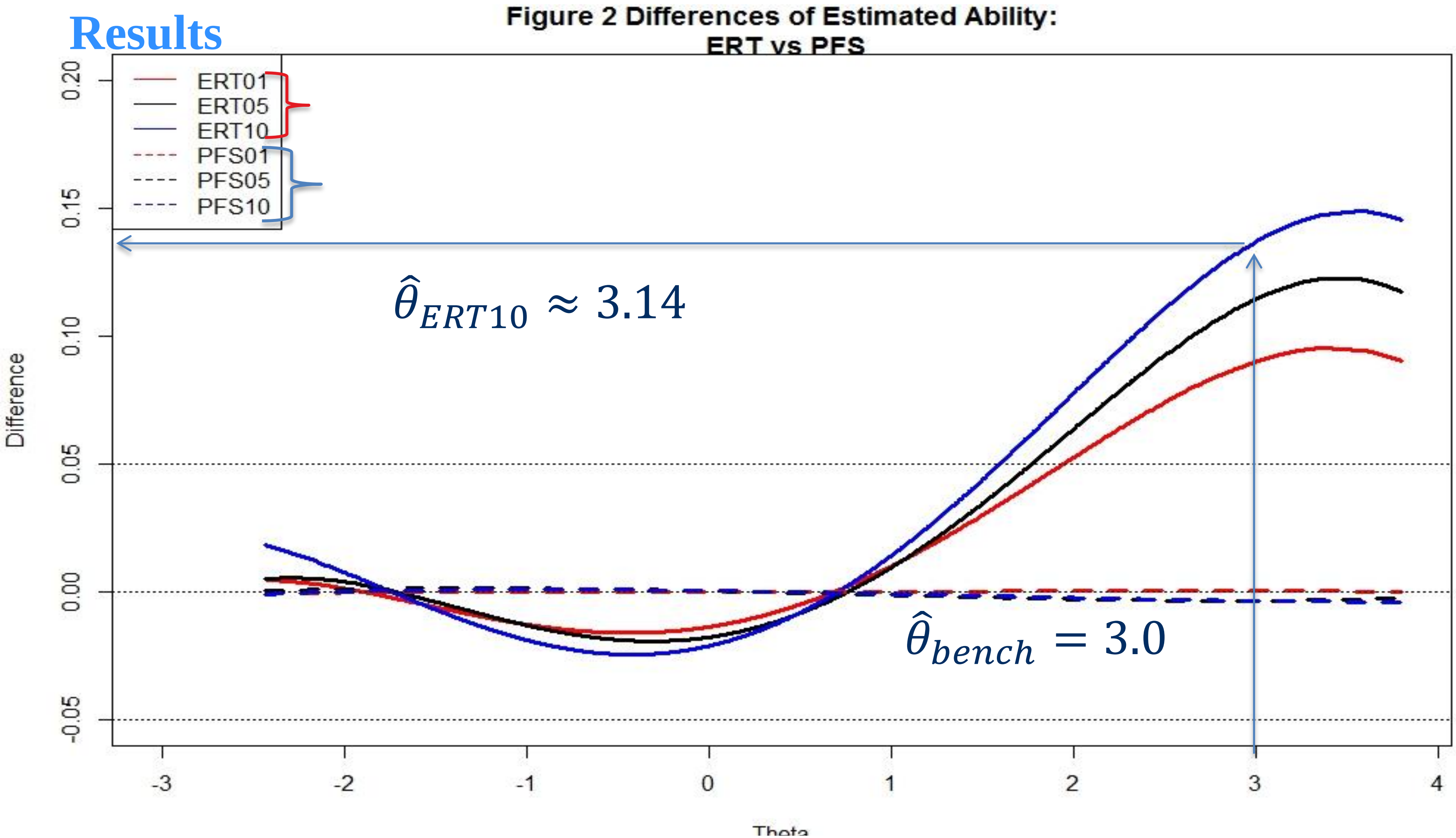
		α=0.01				α=0.05				α=0.10			
Method	***Bench.***	ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT
Item Number	1 , 2 , 4 , 5 , 6 , 7 , 9 , 14 , 15 , 16 , 18 , 23 , 24 , 31 , 34 , 36 , 39 , 40 , 41 , 42 , 51 , 53 , 54	1 , 4 , 6 , 7 , 9 , 12 , 14 , 16 , 18 , 20 , 24 , 34 , 40 , 41 , 45 , 51 , 52 , 53 , 54	1 , 2 , 4 , 5 , 6 , 7 , 9 , 14 , 15 , 16 , 18 , 23 , 24 , 31 , 34 , 36 , 39 , 40 , 41 , 42 , 51 , 53 , 54	1 , 4 , 6 , 7 , 9 , 12 , 14 , 16 , 18 , 20 , 24 , 34 , 40 , 41 , 45 , 51 , 52 , 53 , 54	1 , 4 , 6 , 7 , 9 , 12 , 14 , 16 , 18 , 20 , 24 , 34 , 40 , 41 , 45 , 51 , 52 , 53 , 54	1 , 4 , 6 , 7 , 14 , 15 , 16 , 18 , 20 , 24 , 34 , 37 , 40 , 41 , 43 , 51 , 53 , 54	1 , 2 , 4 , 5 , 6 , 7 , 9 , 14 , 15 , 16 , 18 , 23 , 24 , 31 , 32 , 34 , 36 , 39 , 40 , 41 , 42 , 51 , 53 , 54 ,	1 , 4 , 6 , 7 , 14 , 15 , 16 , 18 , 20 , 24 , 34 , 37 , 40 , 41 , 43 , 51 , 53 , 54	1 , 4 , 6 , 7 , 14 , 15 , 16 , 18 , 20 , 24 , 34 , 37 , 40 , 41 , 43 , 51 , 53 , 54	1 , 6 , 7 , 14 , 15 , 16 , 20 , 23 , 24 , 32 , 34 , 37 , 40 , 41 , 43 , 52 , 53 , 54 , 60	1 , 2 , 4 , 5 , 6 , 7 , 9 , 14 , 15 , 16 , 18 , 23 , 24 , 31 , 32 , 34 , 36 , 39 , 40 , 41 , 42 , 51 , 53 , 54 ,	1 , 6 , 7 , 14 , 15 , 16 , 20 , 24 , 32 , 34 , 37 , 40 , 41 , 43 , 52 , 53 , 54 , 60	1 , 6 , 7 , 14 , 15 , 16 , 20 , 23 , 24 , 32 , 34 , 40 , 41 , 43 , 52 , 53 , 54 , 60
Total	23	19	23	19	19	18	25	18	18	19	25	18	18
N	2734	2372	2734	2372	2372	2218	2729	2216	2211	2118	2728	2112	2112

6 items

3PL

		α=0.01				α=0.05				α=0.10			
Method	***Bench.***	ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT	ERT	PFS	ERT-PFS	PFS-ERT
Item Number	1 , 4 , 6 , 10 , 12 , 15 , 16 , 20 , 23 , 34 , 52 , 53 , 59	1 , 4 , 6 , 7 , 12 , 16 , 20 , 34 , 39 , 52 , 53 , 54	1 , 4 , 6 , 10 , 12 , 15 , 16 , 20 , 34 , 52 , 53 , 59	1 , 4 , 6 , 12 , 16 , 30 , 34 , 42 , 52 , 53	1 , 4 , 6 , 7 , 12 , 16 , 20 , 34 , 39 , 42 , 52 , 53 , 54	1 , 6 , 7 , 12 , 16 , 20 , 30 , 34 , 39 , 42 , 53 , 54	1 , 4 , 6 , 10 , 12 , 15 , 16 , 20 , 23 , 34 , 37 , 52 , 53	1 , 6 , 7 , 12 , 16 , 20 , 30 , 34 , 42 , 53 , 54	1 , 6 , 12 , 20 , 22 , 34 , 42 , 43 , 53 , 54 , 55	1 , 6 , 12 , 16 , 20 , 25 , 34 , 53	1 , 4 , 6 , 10 , 12 , 15 , 16 , 20 , 23 , 34 , 37 , 52 , 53	1 , 4 , 6 , 12 , 16 , 30 , 34 , 42 , 52 , 53	1 , 6 , 16 , 20 , 22 , 34 , 52 , 53
Total	13	12	12	10	13	12	13	11	11	8	13	10	8
N	2734	2376	2728	2373	2372	2228	2714	2220	2218	2125	2691	2102	2088

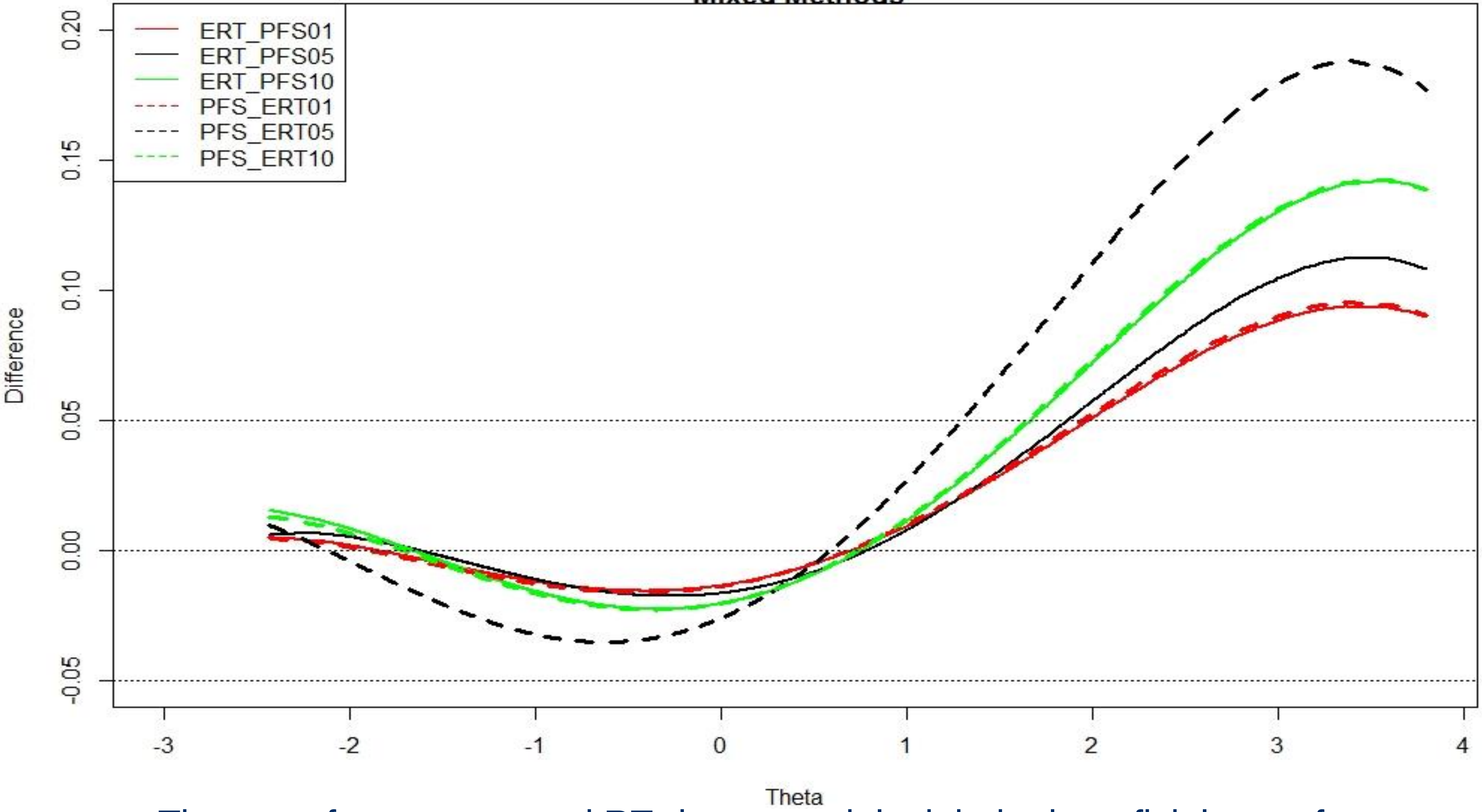
Results



The ERT methods might be beneficial for able students (the theta range of higher than 1.5)

Results

Figure 3 Differences of Estimated Ability: Mixed Methods



The use of responses and RTs into a model might be beneficial more for able students than less able students

Discussion and Conclusions

This study shows that using *ERT* and/or *PFS* for excluding data irregularities would produce a clean data set and might improve the goodness-of-fit, particularly if the 3-PL is implemented.

Comparing *ERT* and *PFS* methods, *ERT* worked better.

Using *PFS* and *ERT* methods together (*PFS_ERT*) might be the best.

The use of responses and RTs into a model for data cleaning might be beneficial more for able students than less able students.

Limitations

- The test is a timed test that might be considered as a difficult test.

Therefore, the results might be different if the test is not timed.

- The item parameter and ability estimations were based on the 2-PL and 3-PL models, which may not be the right model for this test since some of the items were clustered into a testlet.

Therefore, different and/or better results might be obtained if a different model was implemented.

Limitations

- This is not a simulation study in which true abilities of the examinees are unknown.

To know if *ERT* and the mixed methods would produce more accurate estimated abilities for able students as shown in Figures 2 and 3, a simulation study might be required.

Thank you !