Psychometric analysis

Tomasz Żółtak

2025-05-18

1 Data

1.1 Fieldwork timeline

				Number of excluded records				
Day	Number of sessions	Number of all records	Accepted	Overall	Not completed	Speeding	Left the browser tab	Same login used more than one time	Starte to la
09.09	3	54	53	1	0	0	0	1	
13.09	1	24	19	5	2	1	3	0	
20.09	1	11	9	2	1	1	0	0	
23.09	2	33	26	7	4	4	3	0	
24.09	6	86	60	26	26	23	0	0	
25.09	1	25	24	1	0	0	1	0	
26.09	2	48	36	12	2	1	9	0	
27.09	2	39	32	7	2	2	4	0	
30.09	3	66	65	1	0	1	0	0	
02.10	5	64	60	4	2	2	1	1	
03.10	2	27	20	7	2	1	5	0	
04.10	6	87	81	6	1	1	0	5	
08.10	10	148	107	41	12	5	24	0	
09.10	2	20	19	1	1	0	0	0	

				Number of excluded records					
Day	Number of sessions	Number of all records	Accepted	Overall	Not completed	Speeding	Left the browser tab	Same login used more than one time	Starte to la
10.10	1	11	11	0	0	0	0	0	
11.10	1	13	13	0	0	0	0	0	
15.10	1	14	14	0	0	0	0	0	
16.10	4	78	74	4	3	3	1	1	
17.10	1	18	17	1	1	0	0	0	
18.10	4	74	46	28	0	0	0	0	2
21.10	2	33	29	4	4	0	0	0	
22.10	4	39	38	1	0	0	0	0	
23.10	4	44	38	6	1	1	0	4	
24.10	3	34	32	2	0	0	0	2	
25.10	2	24	24	0	0	0	0	0	
28.10	2	18	18	0	0	0	0	0	
29.10	2	47	42	5	1	1	3	0	
30.10	3	40	40	0	0	0	0	0	
05.11	2	41	41	0	0	0	0	0	
06.11	3	50	50	0	0	0	0	0	
07.11	3	47	47	0	0	0	0	0	
14.11	2	40	16	24	12	11	0	0	1
Overall	90	1397	1201	196	77	58	54	14	5

1.2 Distributions of the quota variables

	Number
Overall	1201

Gender	Number	Percent
Male	604	50.3%
Female	597	49.7%

Locality size	Number	Percent
<5k	161	13.4%
5k-100k	770	64.1%
>100k	270	22.5%

1.3 Booklets distribution

Booklet	Number	Percent
1A	80	6.7%
1B	70	5.8%
2A	81	6.7%
2B	75	6.2%
3A	78	6.5%
3B	67	5.6%
4A	85	7.1%
4B	74	6.2%
5A	77	6.4%
5B	71	5.9%
6A	77	6.4%
6B	71	5.9%
$\chi^2(15) = 6.83, p = 0.962$		

Booklet	Number	Percent
7A	68	5.7%
7B	67	5.6%
8A	85	7.1%
8B	75	6.2%
$\chi^2(15) = 6.83, p = 0.962$		

1.4 Items excluded from the analysis

The following items were excluded from the analyses because of having no variance in scores:

- S100006 (only 0)

Also, item **M051134** was excluded because of the technical error which occured during its implementation at the testing platform.

1.5 Drop-off and leaving the test

Number of respondents dropping-off at a given type of test screen

Drop-off frequency

	N	Pct.
Non-item screen	28	37.3%
LLM-generated item	31	41.3%
Item from TIMSS	16	21.3%
Overall	75	100.0%

Frequency of leaving the browser tab (analysise at the test screen level)

Test taker's status	LLM	TIMSS	N	Welch Two Sample t-test
completed	0.00%	0.00%	58,849	no difference at all
excluded	1.15%	2.06%	5,540	t(4922) = -2.7, p = 0.007
not completed	0.22%	0.11%	1,862	t(1615) = 0.61, p = 0.543

2 Model fit

IRT 2PL models were estimated using the *mirt* R package (Chalmers, 2012). Threre were three unidimensional models estimated for each content domain: 1) including all items (i.e. both TIMSS and LLM-generated), 2) including only TIMSS items, 3) including only LLM-generated items. Additionally, for each domain a two-dimensional model was estimated with TIMSS items and LLM-generated items assigned to separate, correlated dimensions.

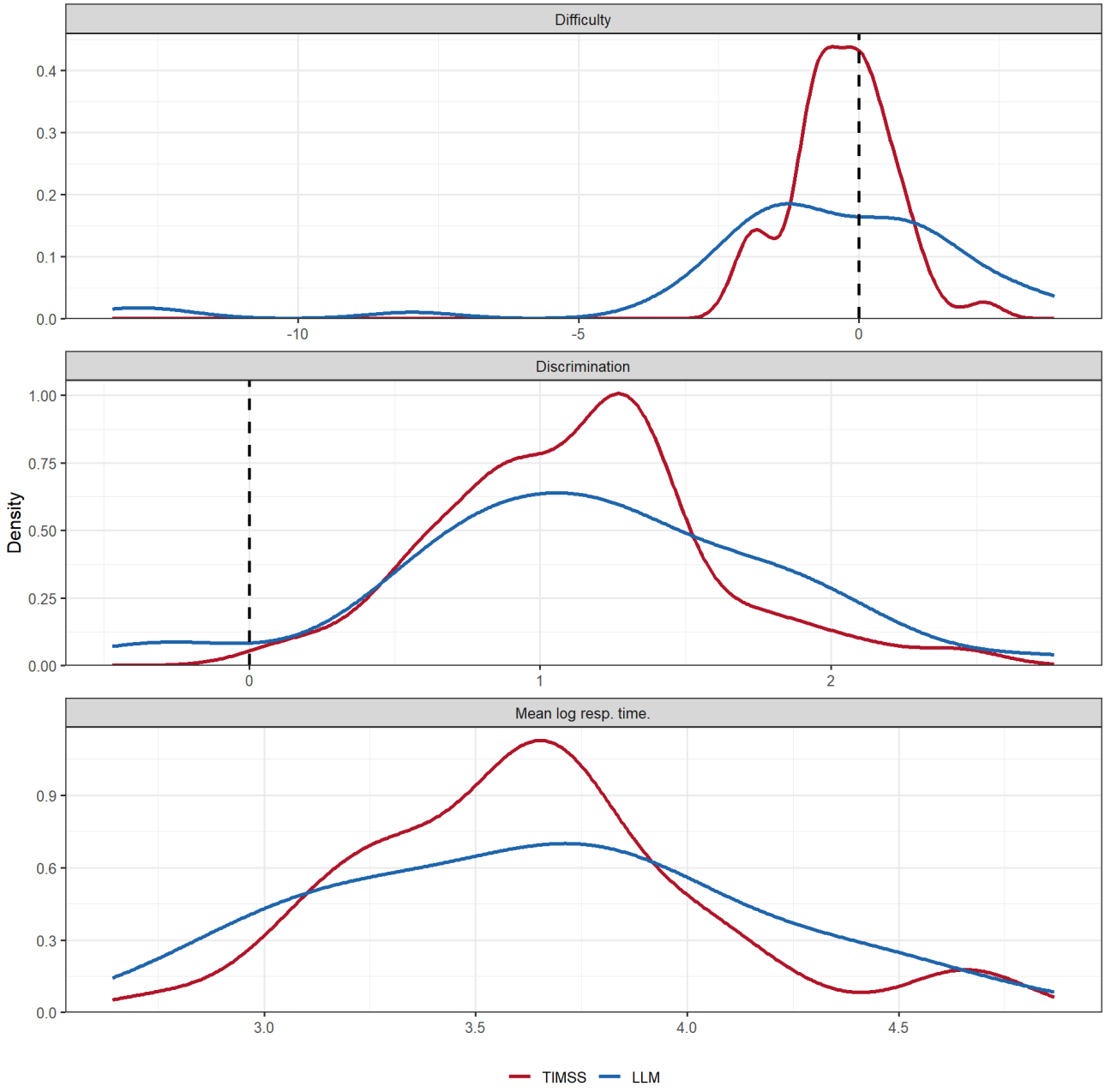
Table below contains values of fit indices for each model, results of the LRT test of the two-dimensional model against the unidimensional model includding all the items and a latent correlation between the TIMMS and LLM-generated dimension in the two-dimensional models.

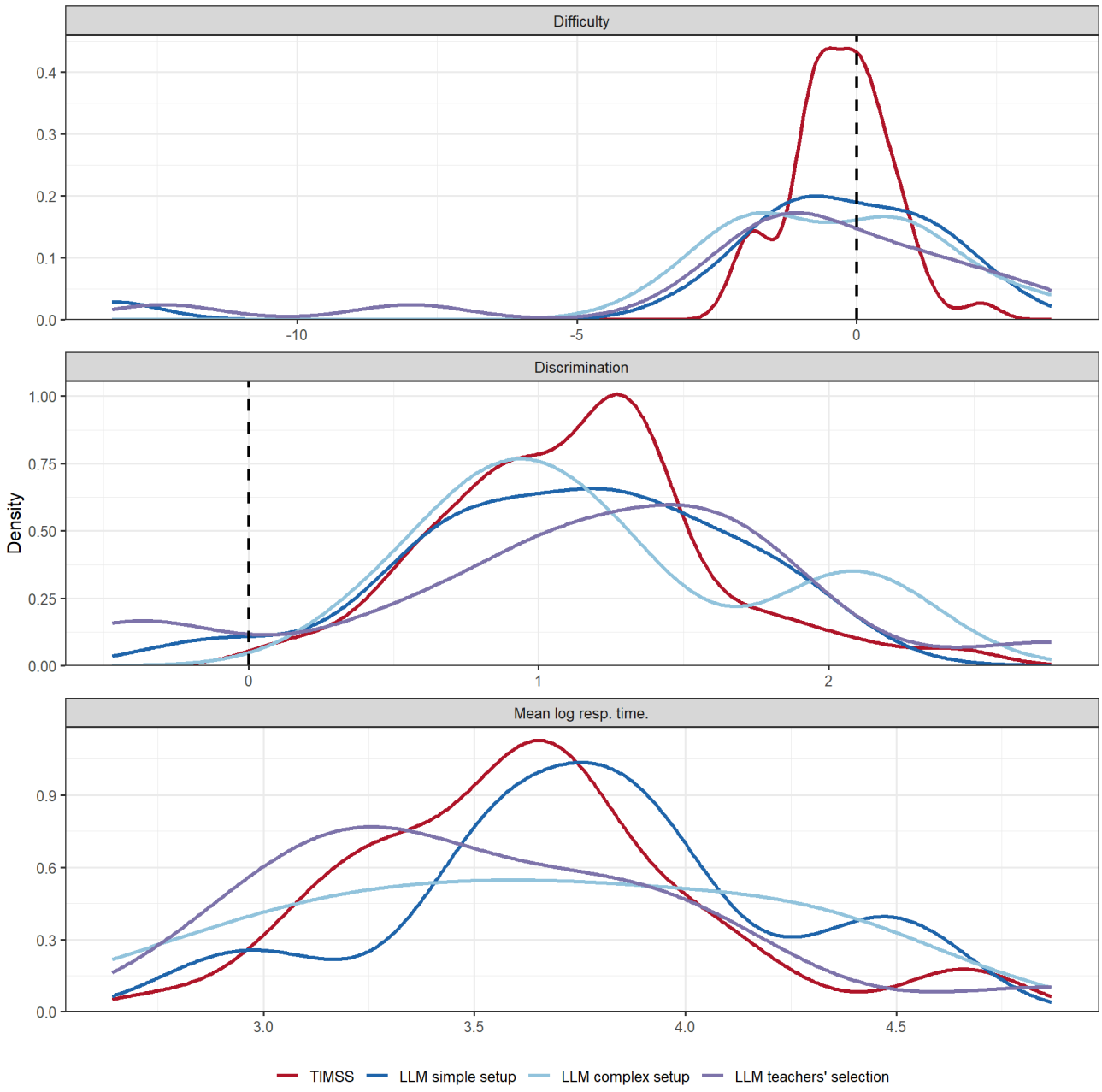
	logLik	AIC	BIC	SABIC	LRT	Latent corr. between dimensions (with 95% CI)
All math items - unidimensional	−14,936.2	30,260.4	31,248.1	30,631.9		
All math items - separate TIMSS and LLM dimensions	−14,914.0	30,218.1	31,210.8	30,591.4	$\chi^2(1) = 44.4$, $p = 0.000$	0.875 (0.833, 0.916)
Only TIMSS math items	−8,373.3	16,936.6	17,420.2	17,118.4		
Only LLM generated math items	−6,837.6	13,873.1	14,377.1	14,062.7		
All science items - unidimensional	−15,064.8	30,523.7	31,526.6	30,900.8		
All science items - separate TIMSS and LLM dimensions	−15,058.2	30,512.3	31,520.3	30,891.4	$\chi^2(1) = 13.4$, $p = 0.000$	0.902 (0.848, 0.955)
Only TIMSS science items	−8,400.0	16,999.9	17,509.0	17,191.4		
Only LLM generated science items	−6,894.5	13,983.0	14,476.9	14,168.7		

3 Distributions of item parameters

3.1 Math items

3.1.1 All items





Source	N	Min	Q1	Median	Q3	Max	Mean	SD
Discrimination								
TIMSS	47	0.10	0.88	1.20	1.35	2.44	1.12	0.46
LLM (all items)	48	−0.47	0.79	1.19	1.59	2.77	1.16	0.64
LLM simple setup	16	−0.10	0.74	1.19	1.42	1.88	1.11	0.53
LLM complex setup	16	0.39	0.80	1.05	1.47	2.27	1.21	0.59
LLM teachers' selection	16	−0.47	0.86	1.24	1.62	2.77	1.15	0.81
Difficulty								
TIMSS	47	−2.13	−0.84	−0.30	0.20	2.23	−0.30	0.89
LLM (all items)	48	−13.31	−1.76	−0.51	0.88	3.49	−0.92	3.22

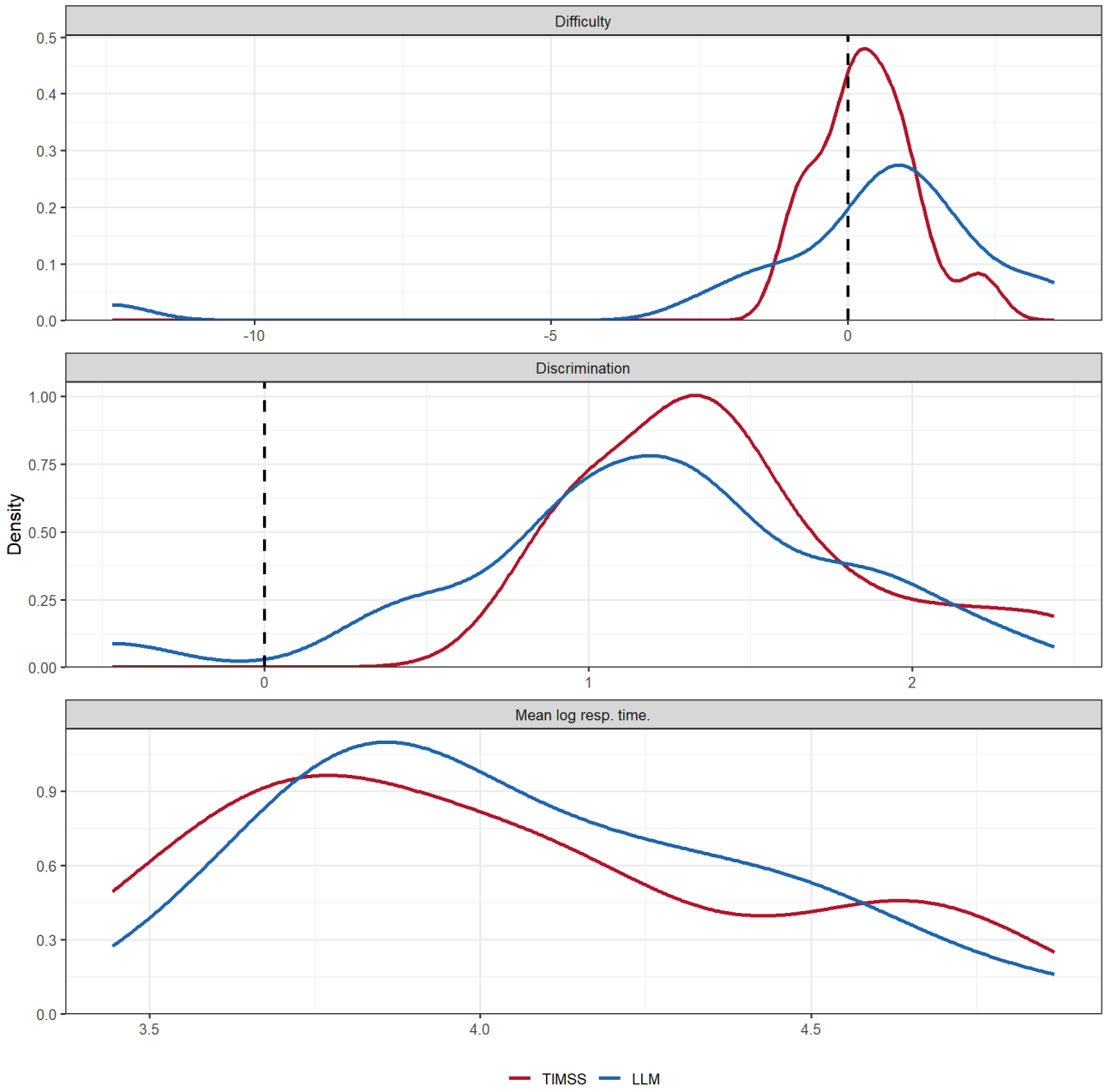
Source	N	Min	Q1	Median	Q3	Max	Mean	SD
LLM simple setup	16	−13.31	−1.38	−0.41	0.88	2.50	−0.97	3.60
LLM complex setup	16	−3.51	−1.85	−0.38	0.76	3.49	−0.37	1.90
LLM teachers' selection	16	−12.40	−1.82	−0.93	0.89	3.37	−1.41	3.91
Mean log resp. time								
TIMSS	47	2.74	3.34	3.62	3.79	4.71	3.63	0.42
LLM (all items)	48	2.64	3.24	3.64	3.98	4.87	3.66	0.52
LLM simple setup	16	2.87	3.55	3.81	3.98	4.54	3.77	0.46
LLM complex setup	16	2.64	3.23	3.67	4.14	4.59	3.65	0.57
LLM teachers' selection	16	2.85	3.20	3.44	3.85	4.87	3.55	0.52

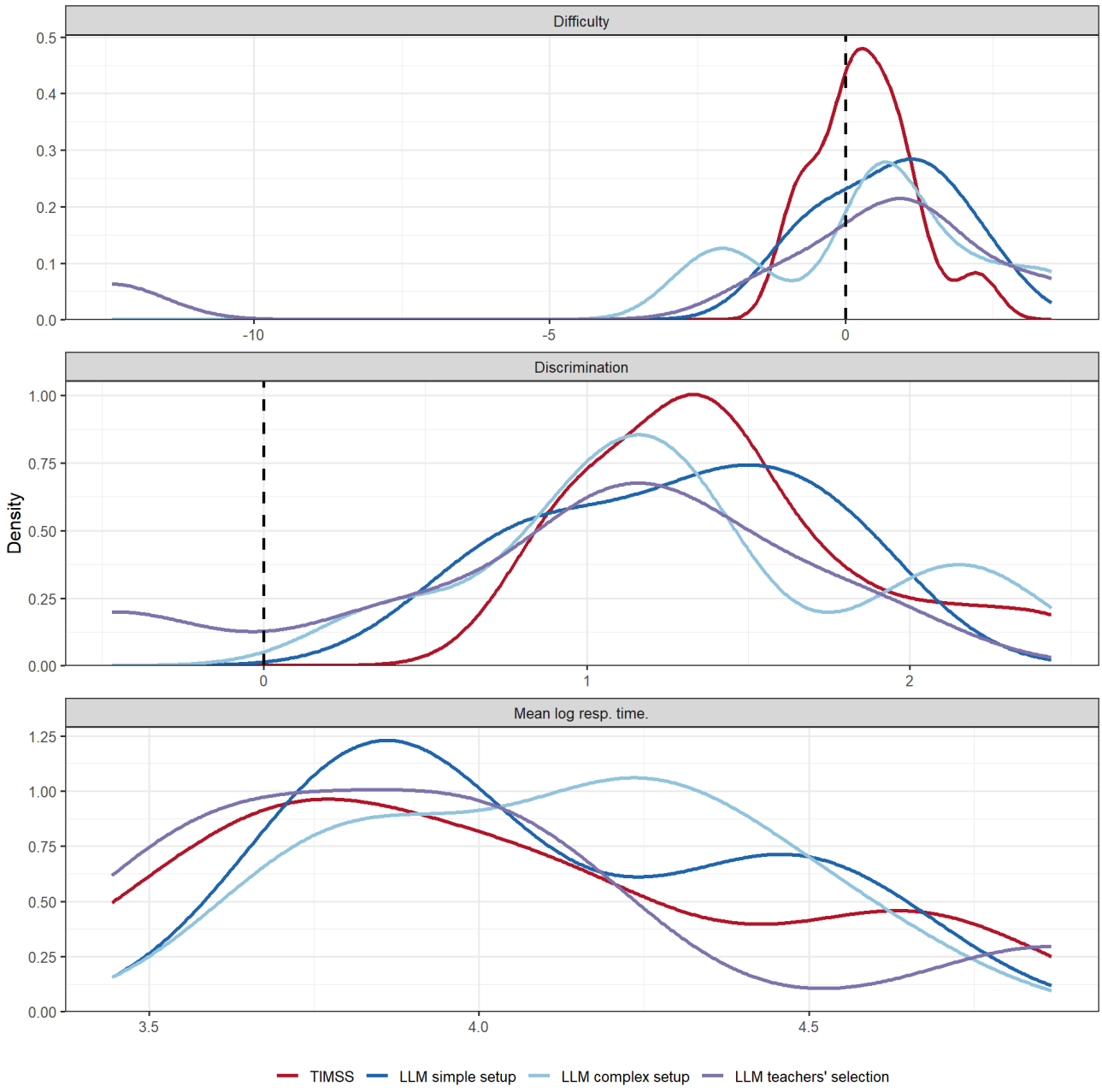
	Kruskal-Wallis rank sum test	ANOVA
Discrimination	$\chi^2(3) = 0.35$, $p = 0.951$	$F(3, 91) = 0.11$, $p = 0.952$
Difficulty	$\chi^2(3) = 0.72$, $p = 0.867$	$F(3, 91) = 1.05$, $p = 0.373$
Mean log resp. time	$\chi^2(3) = 2.53$, $p = 0.471$	$F(3, 91) = 0.65$, $p = 0.588$

Comparison	Wilcoxon rank sum exact test	t-test
Discrimination		
TIMSS vs. LLM (all items)	W = 1081.00, p = 0.730	t(85) = -0.29, p = 0.773
TIMSS vs. LLM simple setup	W = 371.00, p = 0.944	t(22.9) = 0.10, p = 0.923
TIMSS vs. LLM complex setup	W = 377.00, p = 0.994	t(21.6) = -0.54, p = 0.598
TIMSS vs. LLM teachers' selection	W = 333.00, p = 0.506	t(18.4) = -0.13, p = 0.900
LLM simple setup vs. LLM complex setup	W = 122.00, p = 0.838	t(29.7) = -0.51, p = 0.615
LLM simple setup vs. LLM teachers' selection	W = 119.00, p = 0.752	t(26) = -0.17, p = 0.865

Comparison	Wilcoxon rank sum exact test	t-test
LLM complex setup vs. LLM teachers' selection	W = 122.00, p = 0.838	t(27.3) = 0.24, p = 0.815
Difficulty		
TIMSS vs. LLM (all items)	W = 1207.00, p = 0.561	t(54.4) = 1.29, p = 0.204
TIMSS vs. LLM simple setup	W = 382.00, p = 0.932	t(15.6) = 0.74, p = 0.470
TIMSS vs. LLM complex setup	W = 392.00, p = 0.809	t(17.3) = 0.15, p = 0.882
TIMSS vs. LLM teachers' selection	W = 433.00, p = 0.376	t(15.5) = 1.13, p = 0.276
LLM simple setup vs. LLM complex setup	W = 135.00, p = 0.809	t(22.8) = -0.59, p = 0.562
LLM simple setup vs. LLM teachers' selection	W = 142.00, p = 0.616	t(29.8) = 0.33, p = 0.742
LLM complex setup vs. LLM teachers' selection	W = 136.00, p = 0.780	t(21.7) = 0.96, p = 0.349
Mean log resp. time		
TIMSS vs. LLM (all items)	W = 1096.00, p = 0.815	t(89.8) = -0.26, p = 0.797
TIMSS vs. LLM simple setup	W = 299.00, p = 0.230	t(24.1) = -1.10, p = 0.283
TIMSS vs. LLM complex setup	W = 365.00, p = 0.870	t(20.7) = -0.11, p = 0.914
TIMSS vs. LLM teachers' selection	W = 432.00, p = 0.385	t(22) = 0.59, p = 0.560
LLM simple setup vs. LLM complex setup	W = 144.00, p = 0.564	t(28.6) = 0.68, p = 0.500
LLM simple setup vs. LLM teachers' selection	W = 167.00, p = 0.149	t(29.5) = 1.31, p = 0.199
LLM complex setup vs. LLM teachers' selection	W = 146.00, p = 0.515	t(29.7) = 0.53, p = 0.602

3.1.2 Constructed response items





Source	N	Min	Q1	Median	Q3	Max	Mean	SD
Discrimination								
TIMSS	14	0.89	1.13	1.37	1.60	2.44	1.43	0.45
LLM (all items)	23	−0.47	0.94	1.24	1.50	2.27	1.18	0.60
LLM simple setup	8	0.61	0.95	1.35	1.60	1.84	1.29	0.44
LLM complex setup	8	0.39	0.96	1.24	1.49	2.27	1.30	0.61
LLM teachers' selection	7	−0.47	0.70	1.04	1.34	1.88	0.93	0.76
Difficulty								
TIMSS	14	−0.84	−0.06	0.25	0.83	2.23	0.34	0.84
LLM (all items)	23	−12.40	−0.20	0.81	1.58	3.49	0.14	3.11

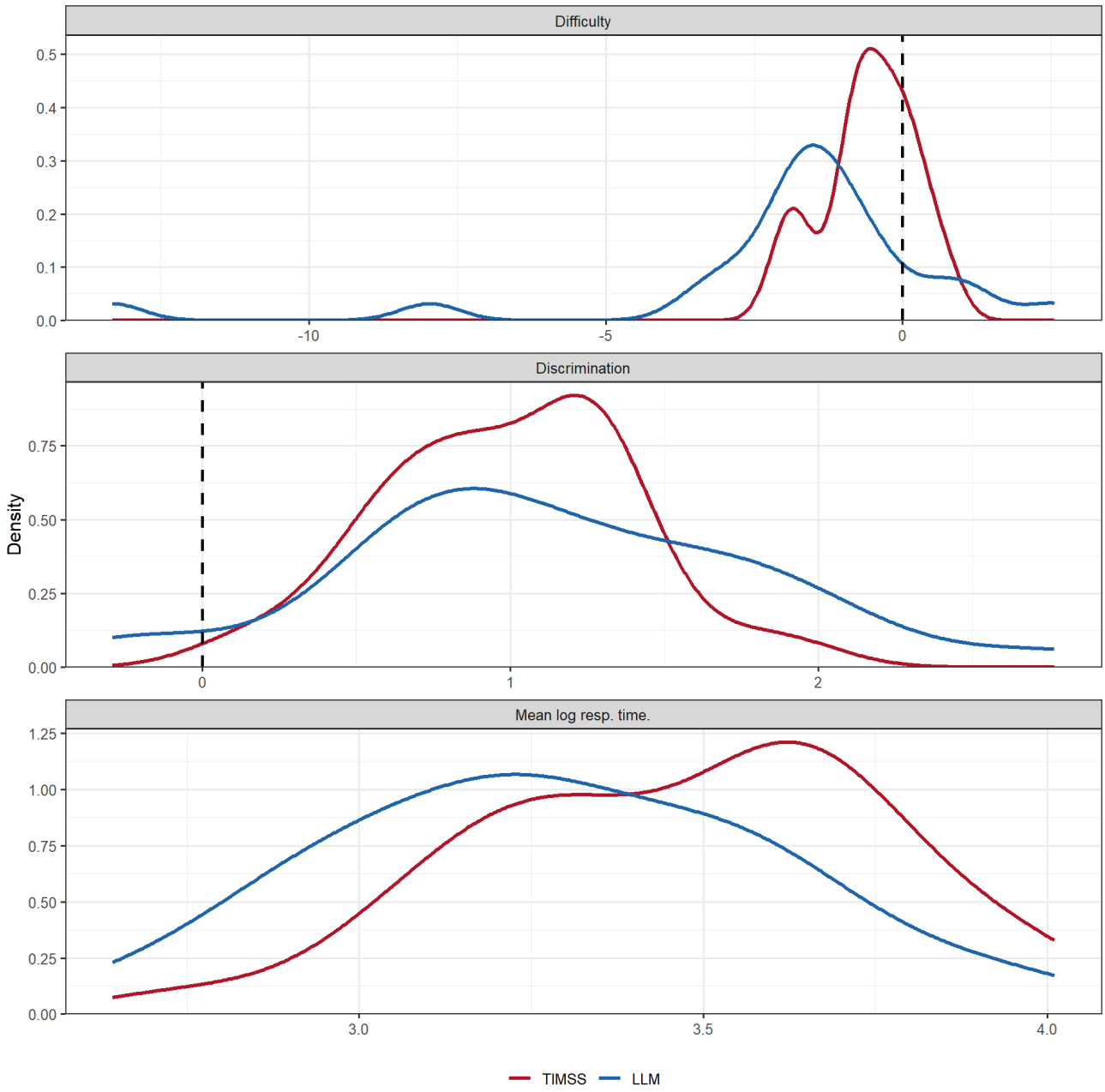
Source	N	Min	Q1	Median	Q3	Max	Mean	SD
LLM simple setup	8	−1.08	−0.11	0.86	1.57	2.50	0.74	1.18
LLM complex setup	8	−2.47	−0.19	0.61	1.34	3.49	0.51	1.92
LLM teachers' selection	7	−12.40	−0.63	0.81	1.36	3.37	−0.96	5.24
Mean log resp. time								
TIMSS	14	3.44	3.73	3.95	4.21	4.71	4.01	0.41
LLM (all items)	23	3.53	3.82	3.98	4.31	4.87	4.06	0.35
LLM simple setup	8	3.78	3.84	3.92	4.38	4.54	4.09	0.33
LLM complex setup	8	3.71	3.85	4.16	4.31	4.59	4.12	0.31
LLM teachers' selection	7	3.53	3.65	3.90	4.08	4.87	3.97	0.45

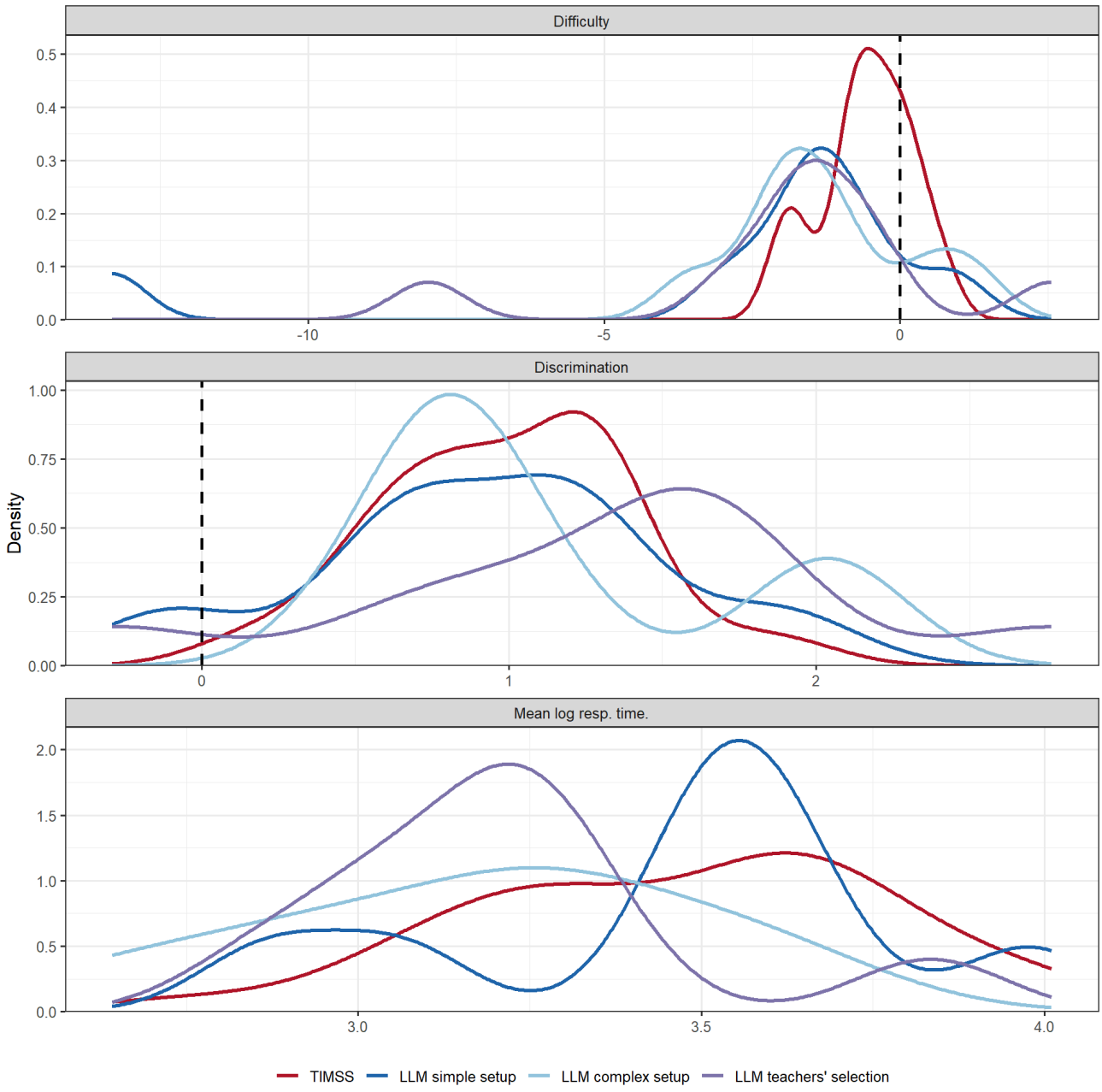
	Kruskal-Wallis rank sum test	ANOVA
Discrimination	$\chi2(3) = 2.12$, $p = 0.548$	$F(3, 33) = 1.28$, $p = 0.298$
Difficulty	$\chi2(3) = 0.79$, $p = 0.851$	$F(3, 33) = 0.68$, $p = 0.573$
Mean log resp. time	$\chi2(3) = 1.38$, $p = 0.709$	$F(3, 33) = 0.26$, $p = 0.855$

Comparison	Wilcoxon rank sum exact test	t-test
Discrimination		
TIMSS vs. LLM (all items)	$W = 199.00$, $p = 0.244$	$t(33.5) = 1.42$, $p = 0.165$
TIMSS vs. LLM simple setup	$W = 64.00$, $p = 0.616$	$t(15) = 0.73$, $p = 0.478$
TIMSS vs. LLM complex setup	$W = 67.00$, $p = 0.482$	$t(11.5) = 0.53$, $p = 0.603$
TIMSS vs. LLM teachers' selection	$W = 68.00$, $p = 0.172$	$t(8.1) = 1.59$, $p = 0.149$
LLM simple setup vs. LLM complex setup	$W = 35.00$, $p = 0.798$	$t(12.8) = -0.05$, $p = 0.965$
LLM simple setup vs. LLM teachers' selection	$W = 34.00$, $p = 0.536$	$t(9.3) = 1.08$, $p = 0.306$

Comparison	Wilcoxon rank sum exact test	t-test
LLM complex setup vs. LLM teachers' selection	W = 34.00, p = 0.536	t(11.5) = 1.02, p = 0.328
Difficulty		
TIMSS vs. LLM (all items)	W = 136.00, p = 0.448	t(27) = 0.28, p = 0.778
TIMSS vs. LLM simple setup	W = 43.00, p = 0.402	t(11.2) = -0.85, p = 0.412
TIMSS vs. LLM complex setup	W = 48.00, p = 0.616	t(8.6) = -0.24, p = 0.813
TIMSS vs. LLM teachers' selection	W = 45.00, p = 0.799	t(6.2) = 0.65, p = 0.537
LLM simple setup vs. LLM complex setup	W = 35.00, p = 0.798	t(11.6) = 0.29, p = 0.779
LLM simple setup vs. LLM teachers' selection	W = 31.00, p = 0.779	t(6.5) = 0.84, p = 0.429
LLM complex setup vs. LLM teachers' selection	W = 28.00, p = 1.000	t(7.4) = 0.71, p = 0.502
Mean log resp. time		
TIMSS vs. LLM (all items)	W = 144.00, p = 0.610	t(24.3) = -0.35, p = 0.727
TIMSS vs. LLM simple setup	W = 48.00, p = 0.616	t(17.7) = -0.45, p = 0.659
TIMSS vs. LLM complex setup	W = 44.00, p = 0.441	t(18.4) = -0.68, p = 0.506
TIMSS vs. LLM teachers' selection	W = 52.00, p = 0.856	t(11.1) = 0.23, p = 0.819
LLM simple setup vs. LLM complex setup	W = 30.00, p = 0.878	t(14) = -0.21, p = 0.835
LLM simple setup vs. LLM teachers' selection	W = 34.00, p = 0.536	t(10.7) = 0.58, p = 0.576
LLM complex setup vs. LLM teachers' selection	W = 38.00, p = 0.281	t(10.3) = 0.75, p = 0.468

3.1.3 Multiple choice items





Source	N	Min	Q1	Median	Q3	Max	Mean	SD
Discrimination								
TIMSS	33	0.10	0.70	0.94	1.25	1.93	0.99	0.40
LLM (all items)	25	−0.29	0.76	1.13	1.60	2.77	1.13	0.68
LLM simple setup	8	−0.10	0.67	0.95	1.24	1.88	0.93	0.59
LLM complex setup	8	0.51	0.79	0.84	1.34	2.11	1.12	0.59
LLM teachers' selection	9	−0.29	0.94	1.59	1.68	2.77	1.32	0.85
Difficulty								
TIMSS	33	−2.13	−0.89	−0.57	0.02	0.82	−0.57	0.78
LLM (all items)	25	−13.31	−2.02	−1.38	−0.57	2.56	−1.90	3.05

Source	N	Min	Q1	Median	Q3	Max	Mean	SD
LLM simple setup	8	−13.31	−2.15	−1.42	−0.85	0.88	−2.69	4.43
LLM complex setup	8	−3.51	−2.02	−1.47	−0.71	1.23	−1.26	1.50
LLM teachers' selection	9	−7.97	−2.02	−1.38	−0.57	2.56	−1.76	2.79
Mean log resp. time								
TIMSS	33	2.74	3.26	3.54	3.66	4.01	3.47	0.30
LLM (all items)	25	2.64	3.04	3.24	3.50	3.98	3.28	0.33
LLM simple setup	8	2.87	3.38	3.53	3.61	3.98	3.46	0.35
LLM complex setup	8	2.64	2.97	3.22	3.36	3.63	3.18	0.32
LLM teachers' selection	9	2.85	3.04	3.20	3.29	3.83	3.22	0.28

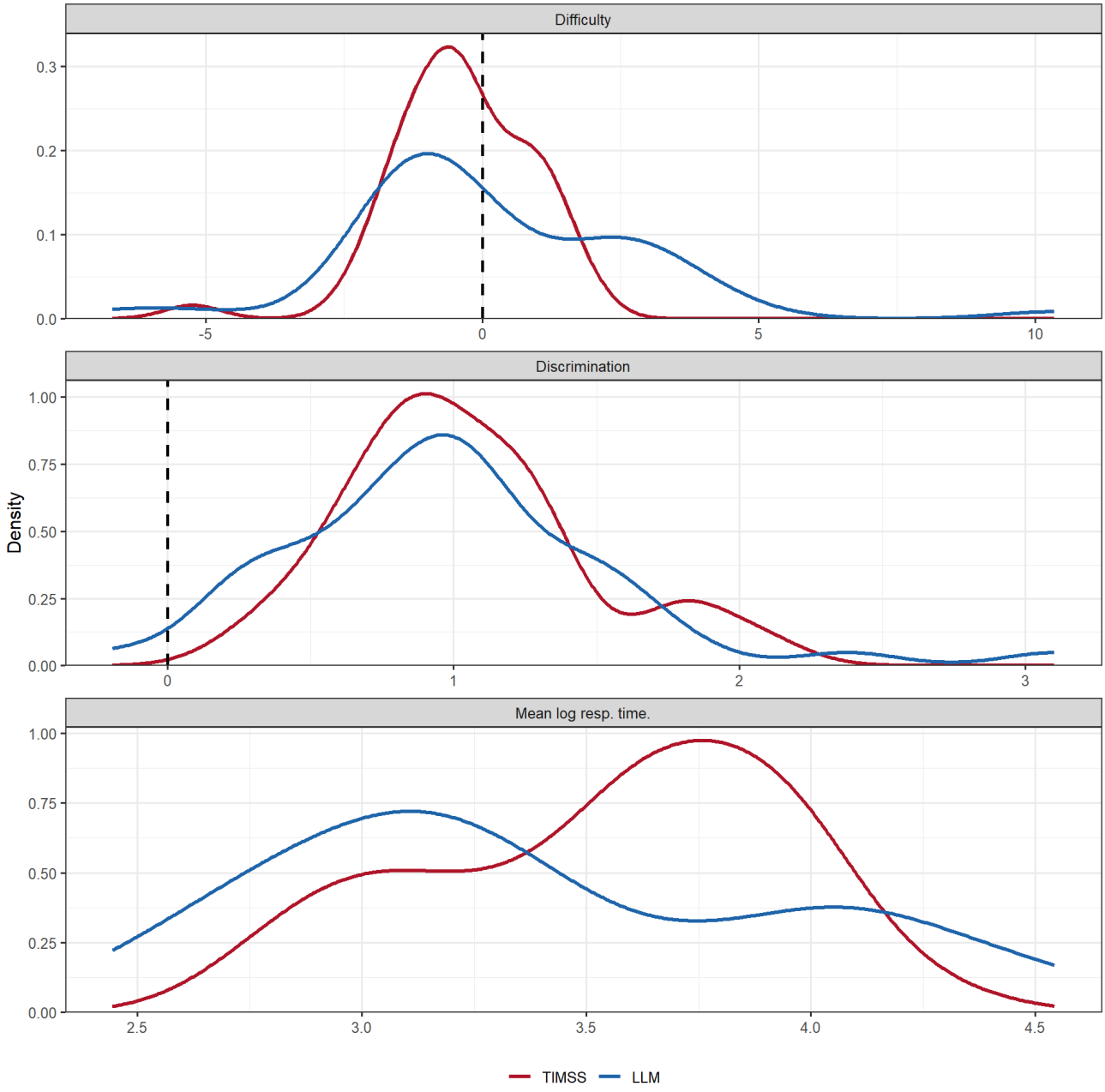
	Kruskal-Wallis rank sum test	ANOVA
Discrimination	$\chi^2(3) = 2.85$, $p = 0.416$	$F(3, 54) = 1.05$, $p = 0.377$
Difficulty	$\chi^2(3) = 7.62$, $p = 0.055$	$F(3, 54) = 2.57$, $p = 0.064$
Mean log resp. time	$\chi^2(3) = 8.53$, $p = 0.036$	$F(3, 54) = 3.07$, $p = 0.035$

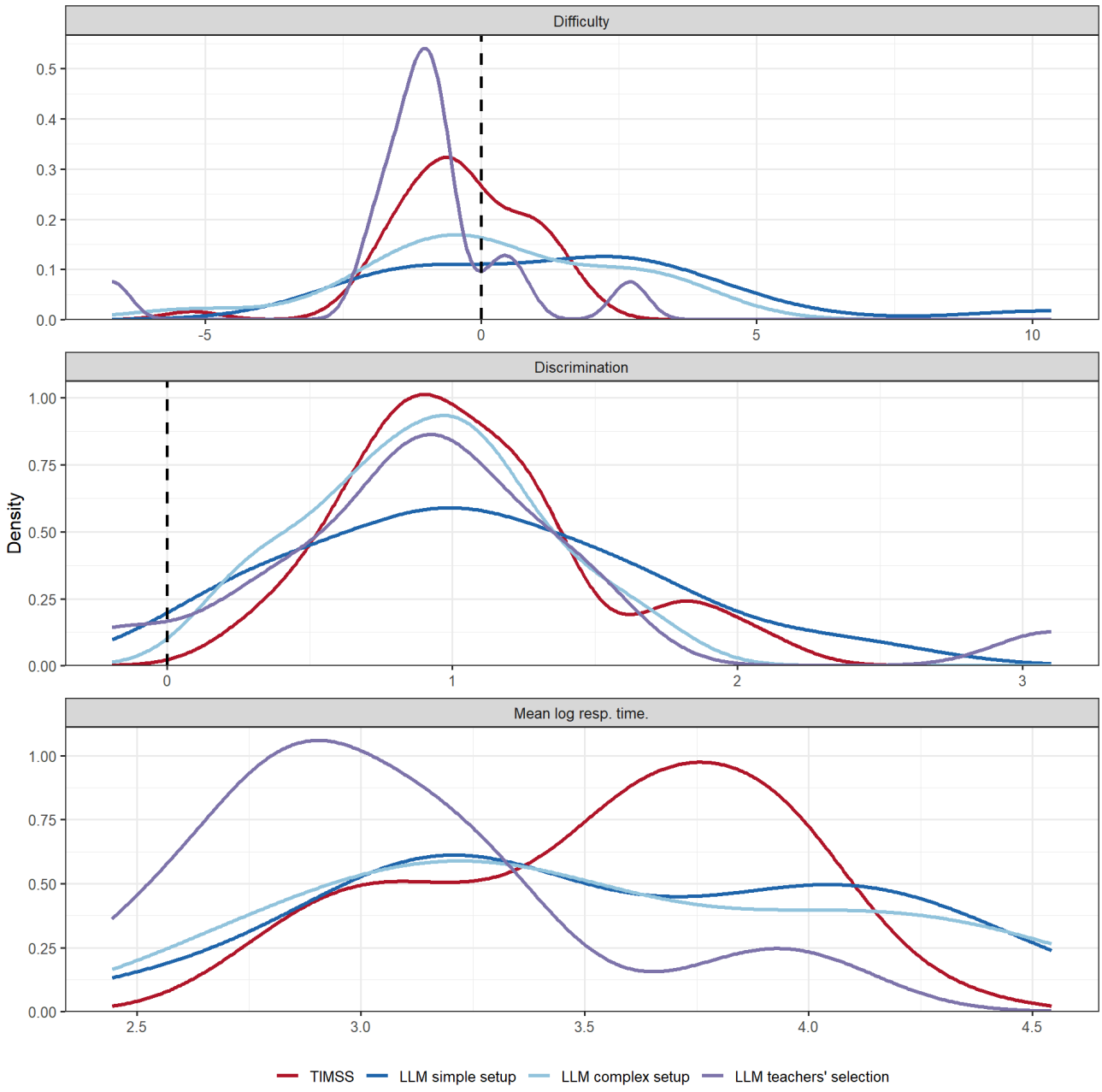
Comparison	Wilcoxon rank sum exact test	t-test
Discrimination		
TIMSS vs. LLM (all items)	W = 371.00, p = 0.523	t(36.2) = -0.90, p = 0.372
TIMSS vs. LLM simple setup	W = 143.00, p = 0.735	t(8.6) = 0.28, p = 0.784
TIMSS vs. LLM complex setup	W = 135.00, p = 0.936	t(8.6) = -0.58, p = 0.578
TIMSS vs. LLM teachers' selection	W = 93.00, p = 0.092	t(9) = -1.13, p = 0.289
LLM simple setup vs. LLM complex setup	W = 27.00, p = 0.645	t(14) = -0.64, p = 0.531
LLM simple setup vs. LLM teachers' selection	W = 23.00, p = 0.236	t(14.2) = -1.11, p = 0.285

Comparison	Wilcoxon rank sum exact test	t-test
LLM complex setup vs. LLM teachers' selection	W = 28.00, p = 0.481	t(14.3) = -0.57, p = 0.578
Difficulty		
TIMSS vs. LLM (all items)	W = 588.00, p = 0.005	t(26.4) = 2.12, p = 0.043
TIMSS vs. LLM simple setup	W = 191.00, p = 0.053	t(7.1) = 1.35, p = 0.220
TIMSS vs. LLM complex setup	W = 184.00, p = 0.091	t(7.9) = 1.26, p = 0.243
TIMSS vs. LLM teachers' selection	W = 213.00, p = 0.049	t(8.3) = 1.27, p = 0.238
LLM simple setup vs. LLM complex setup	W = 31.00, p = 0.959	t(8.6) = -0.86, p = 0.412
LLM simple setup vs. LLM teachers' selection	W = 37.00, p = 0.963	t(11.5) = -0.51, p = 0.622
LLM complex setup vs. LLM teachers' selection	W = 37.00, p = 0.963	t(12.5) = 0.47, p = 0.646
Mean log resp. time		
TIMSS vs. LLM (all items)	W = 559.00, p = 0.021	t(49.4) = 2.21, p = 0.032
TIMSS vs. LLM simple setup	W = 140.00, p = 0.809	t(9.7) = 0.05, p = 0.958
TIMSS vs. LLM complex setup	W = 198.00, p = 0.029	t(10.1) = 2.31, p = 0.043
TIMSS vs. LLM teachers' selection	W = 221.00, p = 0.025	t(13.6) = 2.35, p = 0.035
LLM simple setup vs. LLM complex setup	W = 47.00, p = 0.130	t(13.9) = 1.69, p = 0.112
LLM simple setup vs. LLM teachers' selection	W = 53.00, p = 0.114	t(13.4) = 1.58, p = 0.138
LLM complex setup vs. LLM teachers' selection	W = 36.00, p = 1.000	t(13.9) = -0.29, p = 0.778

3.2 Science items

3.2.1 All items





Source	N	Min	Q1	Median	Q3	Max	Mean	SD
Discrimination								
TIMSS	48	0.23	0.76	0.95	1.25	2.12	1.04	0.42
LLM (all items)	47	−0.19	0.65	0.95	1.22	3.10	0.98	0.58
LLM simple setup	15	0.16	0.66	1.01	1.50	2.38	1.07	0.61
LLM complex setup	16	0.23	0.65	0.90	1.13	1.67	0.91	0.40
LLM teachers' selection	16	−0.19	0.69	0.93	1.20	3.10	0.98	0.72
Difficulty								
TIMSS	48	−5.24	−1.17	−0.42	0.51	1.82	−0.38	1.27
LLM (all items)	47	−6.68	−1.26	−0.63	2.06	10.35	0.21	2.71

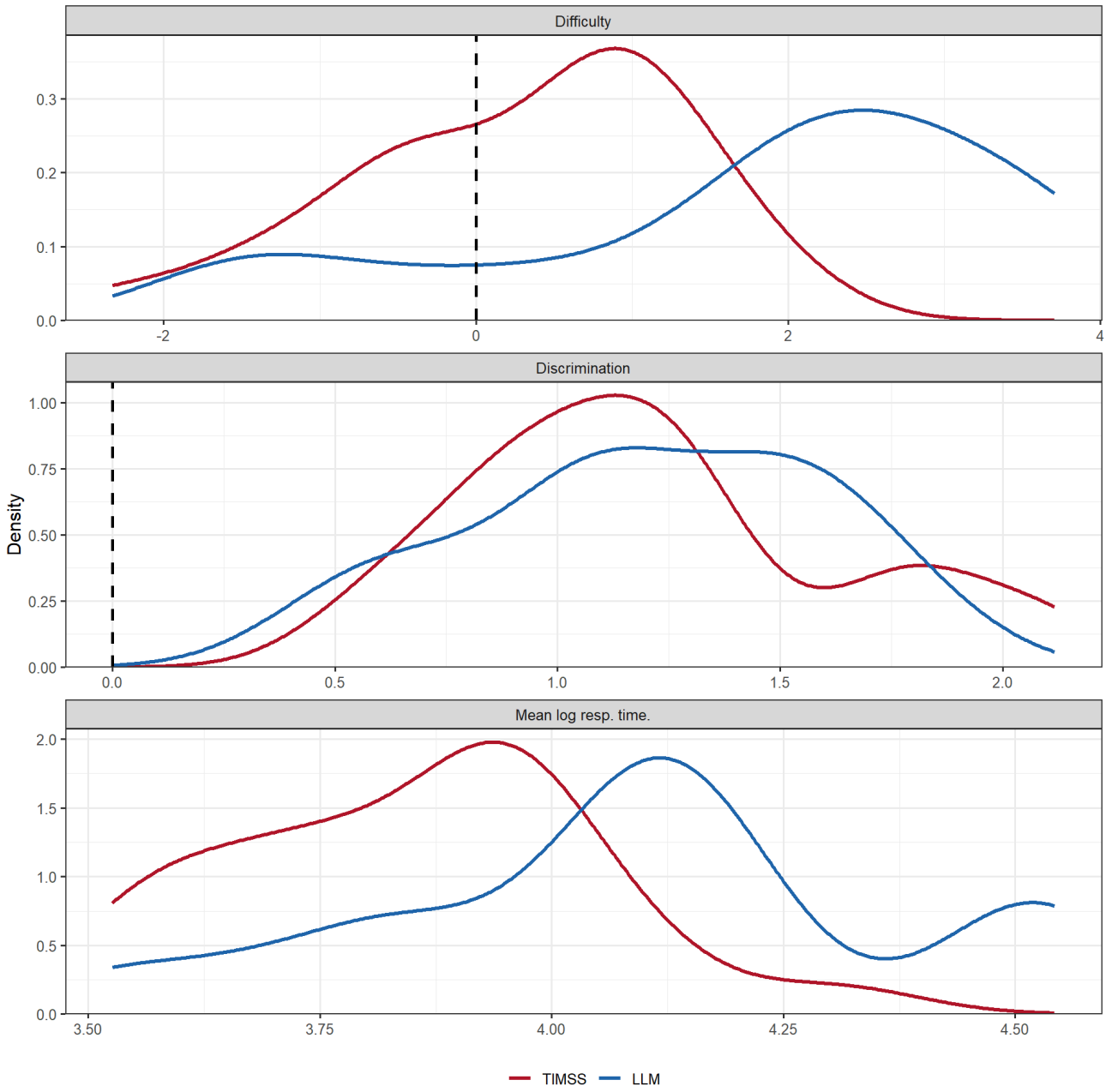
Source	N	Min	Q1	Median	Q3	Max	Mean	SD
LLM simple setup	15	−2.29	−1.18	1.87	2.76	10.35	1.63	3.22
LLM complex setup	16	−5.18	−1.16	0.12	2.18	3.63	0.17	2.33
LLM teachers' selection	16	−6.68	−1.50	−1.02	−0.65	2.70	−1.09	1.88
Mean log resp. time								
TIMSS	48	2.74	3.26	3.62	3.87	4.31	3.55	0.40
LLM (all items)	47	2.44	2.93	3.21	3.86	4.54	3.38	0.58
LLM simple setup	15	2.53	3.17	3.38	4.08	4.52	3.56	0.57
LLM complex setup	16	2.68	3.11	3.42	4.11	4.54	3.54	0.60
LLM teachers' selection	16	2.44	2.77	2.95	3.23	4.03	3.06	0.42

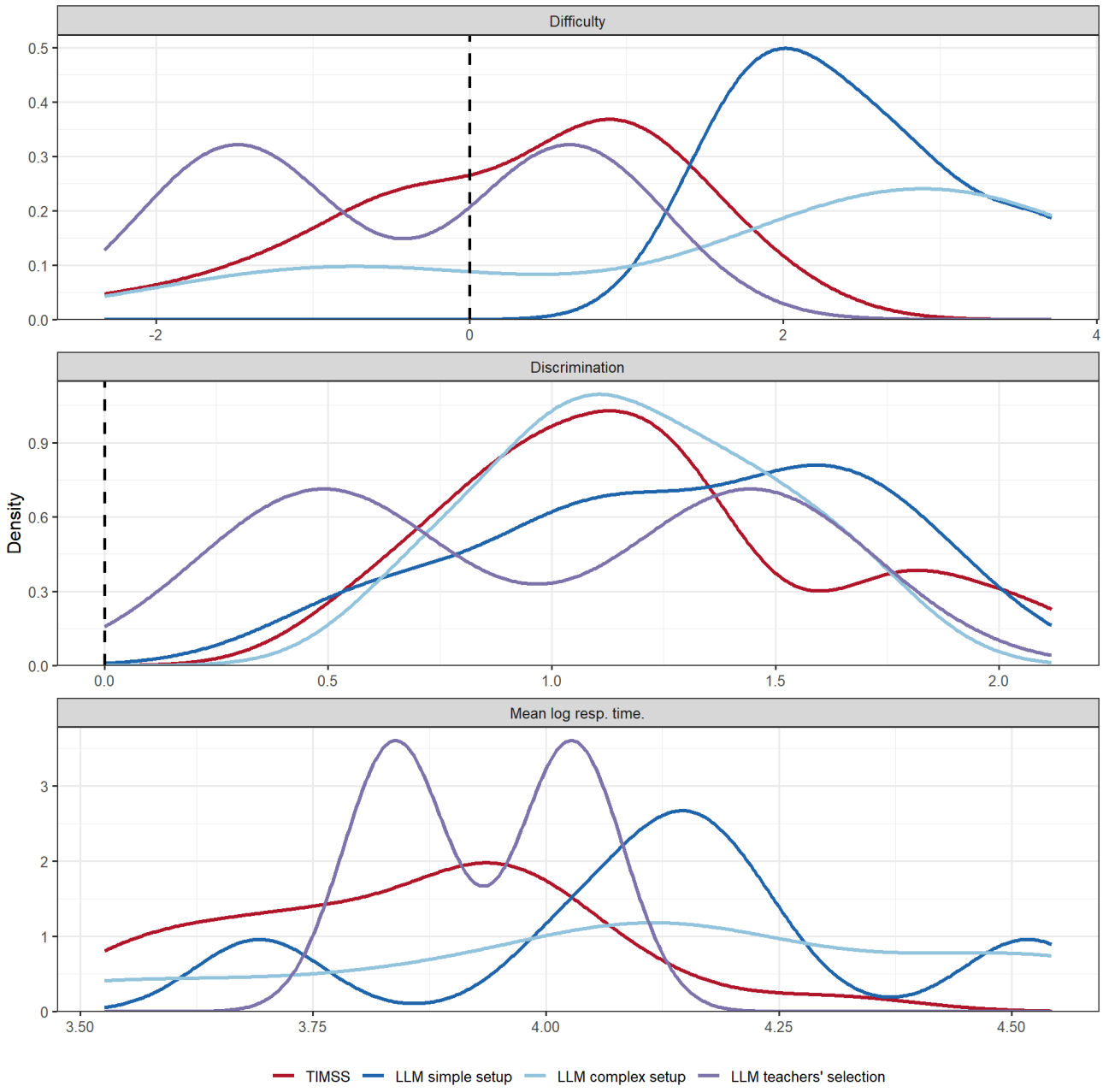
	Kruskal-Wallis rank sum test	ANOVA
Discrimination	$\chi^2(3) = 1.05$, $p = 0.789$	$F(3, 91) = 0.35$, $p = 0.788$
Difficulty	$\chi^2(3) = 9.21$, $p = 0.027$	$F(3, 91) = 5.61$, $p = 0.001$
Mean log resp. time	$\chi^2(3) = 12.41$, $p = 0.006$	$F(3, 91) = 4.78$, $p = 0.004$

Comparison	Wilcoxon rnak sum exact test	t-test
Discrimination		
TIMSS vs. LLM (all items)	W = 1222.00, p = 0.488	t(84) = 0.53, p = 0.598
TIMSS vs. LLM simple setup	W = 354.00, p = 0.930	t(18.4) = -0.17, p = 0.864
TIMSS vs. LLM complex setup	W = 443.00, p = 0.368	t(27.1) = 1.13, p = 0.267
TIMSS vs. LLM teachers' selection	W = 425.00, p = 0.534	t(18.6) = 0.30, p = 0.765
LLM simple setup vs. LLM complex setup	W = 135.00, p = 0.572	t(23.9) = 0.87, p = 0.394
LLM simple setup vs. LLM teachers' selection	W = 134.00, p = 0.599	t(28.7) = 0.36, p = 0.719

Comparison	Wilcoxon rnak sum exact test	t-test
LLM complex setup vs. LLM teachers' selection	W = 127.00, p = 0.985	t(23.5) = -0.37, p = 0.718
Difficulty		
TIMSS vs. LLM (all items)	W = 1055.00, p = 0.591	t(65.1) = -1.35, p = 0.181
TIMSS vs. LLM simple setup	W = 217.00, p = 0.020	t(15.4) = -2.37, p = 0.031
TIMSS vs. LLM complex setup	W = 326.00, p = 0.376	t(18.1) = -0.90, p = 0.378
TIMSS vs. LLM teachers' selection	W = 512.00, p = 0.047	t(19.8) = 1.41, p = 0.173
LLM simple setup vs. LLM complex setup	W = 142.00, p = 0.401	t(25.4) = 1.44, p = 0.162
LLM simple setup vs. LLM teachers' selection	W = 175.00, p = 0.030	t(22.2) = 2.86, p = 0.009
LLM complex setup vs. LLM teachers' selection	W = 173.00, p = 0.094	t(28.7) = 1.69, p = 0.102
Mean log resp. time		
TIMSS vs. LLM (all items)	W = 1352.00, p = 0.096	t(81.8) = 1.61, p = 0.111
TIMSS vs. LLM simple setup	W = 346.00, p = 0.829	t(18.5) = -0.08, p = 0.938
TIMSS vs. LLM complex setup	W = 398.00, p = 0.836	t(19.6) = 0.04, p = 0.970
TIMSS vs. LLM teachers' selection	W = 608.00, p = 0.000	t(24.5) = 4.04, p = 0.000
LLM simple setup vs. LLM complex setup	W = 125.00, p = 0.861	t(29) = 0.09, p = 0.930
LLM simple setup vs. LLM teachers' selection	W = 185.00, p = 0.009	t(25.8) = 2.76, p = 0.011
LLM complex setup vs. LLM teachers' selection	W = 192.00, p = 0.015	t(26.9) = 2.61, p = 0.014

3.2.2 Constructed response items





Source	N	Min	Q1	Median	Q3	Max	Mean	SD
Discrimination								
TIMSS	23	0.53	0.93	1.15	1.36	2.12	1.21	0.43
LLM (all items)	15	0.49	1.01	1.13	1.50	1.82	1.20	0.40
LLM simple setup	6	0.61	1.09	1.34	1.58	1.82	1.30	0.44
LLM complex setup	7	0.71	1.01	1.13	1.38	1.67	1.19	0.32
LLM teachers' selection	2	0.49	0.73	0.97	1.21	1.45	0.97	0.68
Difficulty								
TIMSS	23	−2.33	−0.42	0.43	0.98	1.82	0.26	1.08
LLM (all items)	15	−1.49	1.11	2.07	2.84	3.71	1.78	1.68

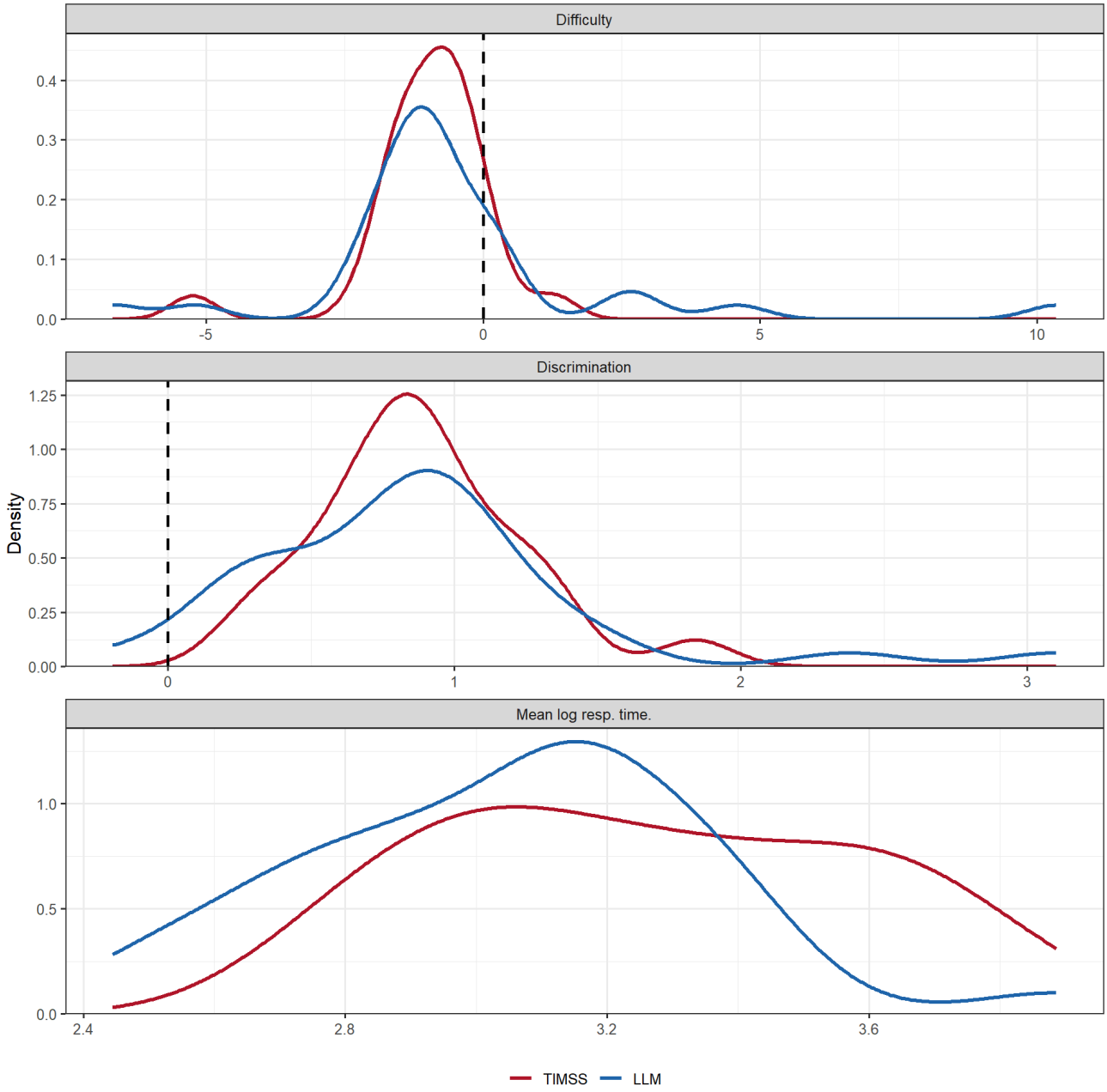
Source	N	Min	Q1	Median	Q3	Max	Mean	SD
LLM simple setup	6	1.57	1.92	2.28	2.80	3.71	2.43	0.78
LLM complex setup	7	−1.38	0.91	2.53	3.13	3.63	1.84	1.92
LLM teachers' selection	2	−1.49	−0.95	−0.42	0.11	0.64	−0.42	1.50
Mean log resp. time								
TIMSS	23	3.55	3.72	3.88	3.97	4.31	3.85	0.20
LLM (all items)	15	3.53	3.93	4.12	4.18	4.54	4.09	0.29
LLM simple setup	6	3.69	4.05	4.14	4.20	4.52	4.12	0.27
LLM complex setup	7	3.53	3.98	4.12	4.32	4.54	4.11	0.35
LLM teachers' selection	2	3.84	3.89	3.93	3.98	4.03	3.93	0.13

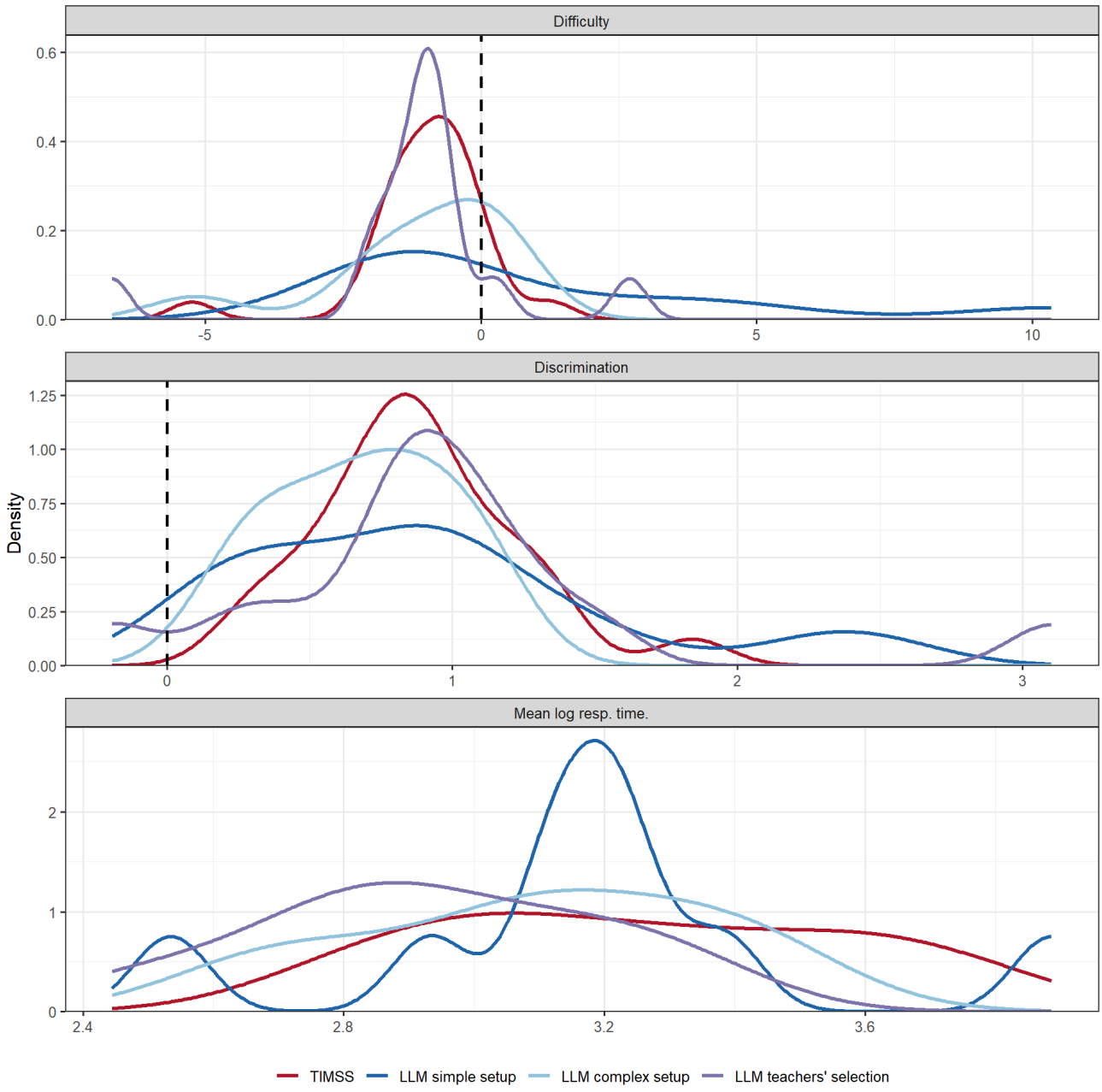
	Kruskal-Wallis rank sum test	ANOVA
Discrimination	$\chi2(3) = 0.72$, $p = 0.869$	$F(3, 34) = 0.31$, $p = 0.818$
Difficulty	$\chi2(3) = 15.44$, $p = 0.001$	$F(3, 34) = 7.08$, $p = 0.001$
Mean log resp. time	$\chi2(3) = 8.61$, $p = 0.035$	$F(3, 34) = 3.38$, $p = 0.029$

Comparison	Wilcoxon rnak sum exact test	t-test
Discrimination		
TIMSS vs. LLM (all items)	$W = 170.00$, $p = 0.953$	$t(31.7) = 0.09$, $p = 0.929$
TIMSS vs. LLM simple setup	$W = 61.00$, $p = 0.694$	$t(7.7) = -0.42$, $p = 0.688$
TIMSS vs. LLM complex setup	$W = 81.00$, $p = 1.000$	$t(13.2) = 0.18$, $p = 0.857$
TIMSS vs. LLM teachers' selection	$W = 28.00$, $p = 0.667$	$t(1.1) = 0.50$, $p = 0.699$
LLM simple setup vs. LLM complex setup	$W = 25.00$, $p = 0.628$	$t(9.1) = 0.51$, $p = 0.620$
LLM simple setup vs. LLM teachers' selection	$W = 9.00$, $p = 0.429$	$t(1.3) = 0.64$, $p = 0.616$

Comparison	Wilcoxon rnak sum exact test	t-test
LLM complex setup vs. LLM teachers' selection	W = 9.00, p = 0.667	t(1.1) = 0.44, p = 0.729
Difficulty		
TIMSS vs. LLM (all items)	W = 71.00, p = 0.002	t(21.5) = -3.09, p = 0.005
TIMSS vs. LLM simple setup	W = 2.00, p = 0.000	t(10.6) = -5.56, p = 0.000
TIMSS vs. LLM complex setup	W = 36.00, p = 0.029	t(7.2) = -2.08, p = 0.075
TIMSS vs. LLM teachers' selection	W = 33.00, p = 0.373	t(1.1) = 0.63, p = 0.635
LLM simple setup vs. LLM complex setup	W = 21.00, p = 1.000	t(8.2) = 0.75, p = 0.472
LLM simple setup vs. LLM teachers' selection	W = 12.00, p = 0.071	t(1.2) = 2.57, p = 0.203
LLM complex setup vs. LLM teachers' selection	W = 12.00, p = 0.222	t(2.1) = 1.76, p = 0.217
Mean log resp. time		
TIMSS vs. LLM (all items)	W = 79.00, p = 0.004	t(22.2) = -2.82, p = 0.010
TIMSS vs. LLM simple setup	W = 24.00, p = 0.014	t(6.5) = -2.32, p = 0.057
TIMSS vs. LLM complex setup	W = 39.00, p = 0.042	t(7.1) = -1.87, p = 0.103
TIMSS vs. LLM teachers' selection	W = 16.00, p = 0.540	t(1.4) = -0.80, p = 0.539
LLM simple setup vs. LLM complex setup	W = 22.00, p = 0.945	t(10.9) = 0.05, p = 0.959
LLM simple setup vs. LLM teachers' selection	W = 9.00, p = 0.429	t(4) = 1.30, p = 0.263
LLM complex setup vs. LLM teachers' selection	W = 11.00, p = 0.333	t(5.4) = 1.09, p = 0.320

3.2.3 Multiple choice items





Source	N	Min	Q1	Median	Q3	Max	Mean	SD
Discrimination								
TIMSS	25	0.23	0.68	0.85	1.05	1.84	0.88	0.36
LLM (all items)	32	−0.19	0.44	0.86	1.04	3.10	0.88	0.63
LLM simple setup	9	0.16	0.36	0.91	1.01	2.38	0.91	0.69
LLM complex setup	9	0.23	0.45	0.65	0.86	1.12	0.69	0.32
LLM teachers' selection	14	−0.19	0.76	0.93	1.14	3.10	0.98	0.75
Difficulty								
TIMSS	25	−5.24	−1.54	−0.82	−0.39	1.29	−0.97	1.16
LLM (all items)	32	−6.68	−1.51	−1.02	0.09	10.35	−0.53	2.80

Source	N	Min	Q1	Median	Q3	Max	Mean	SD
LLM simple setup	9	−2.29	−1.22	−1.18	2.62	10.35	1.10	4.12
LLM complex setup	9	−5.18	−1.82	−0.81	0.17	0.32	−1.13	1.76
LLM teachers' selection	14	−6.68	−1.48	−1.02	−0.70	2.70	−1.19	1.95
Mean log resp. time								
TIMSS	25	2.74	3.02	3.28	3.54	3.87	3.26	0.33
LLM (all items)	32	2.44	2.84	3.11	3.22	3.89	3.05	0.31
LLM simple setup	9	2.53	3.11	3.20	3.25	3.89	3.18	0.36
LLM complex setup	9	2.68	2.94	3.11	3.33	3.42	3.09	0.28
LLM teachers' selection	14	2.44	2.76	2.91	3.13	3.36	2.93	0.27

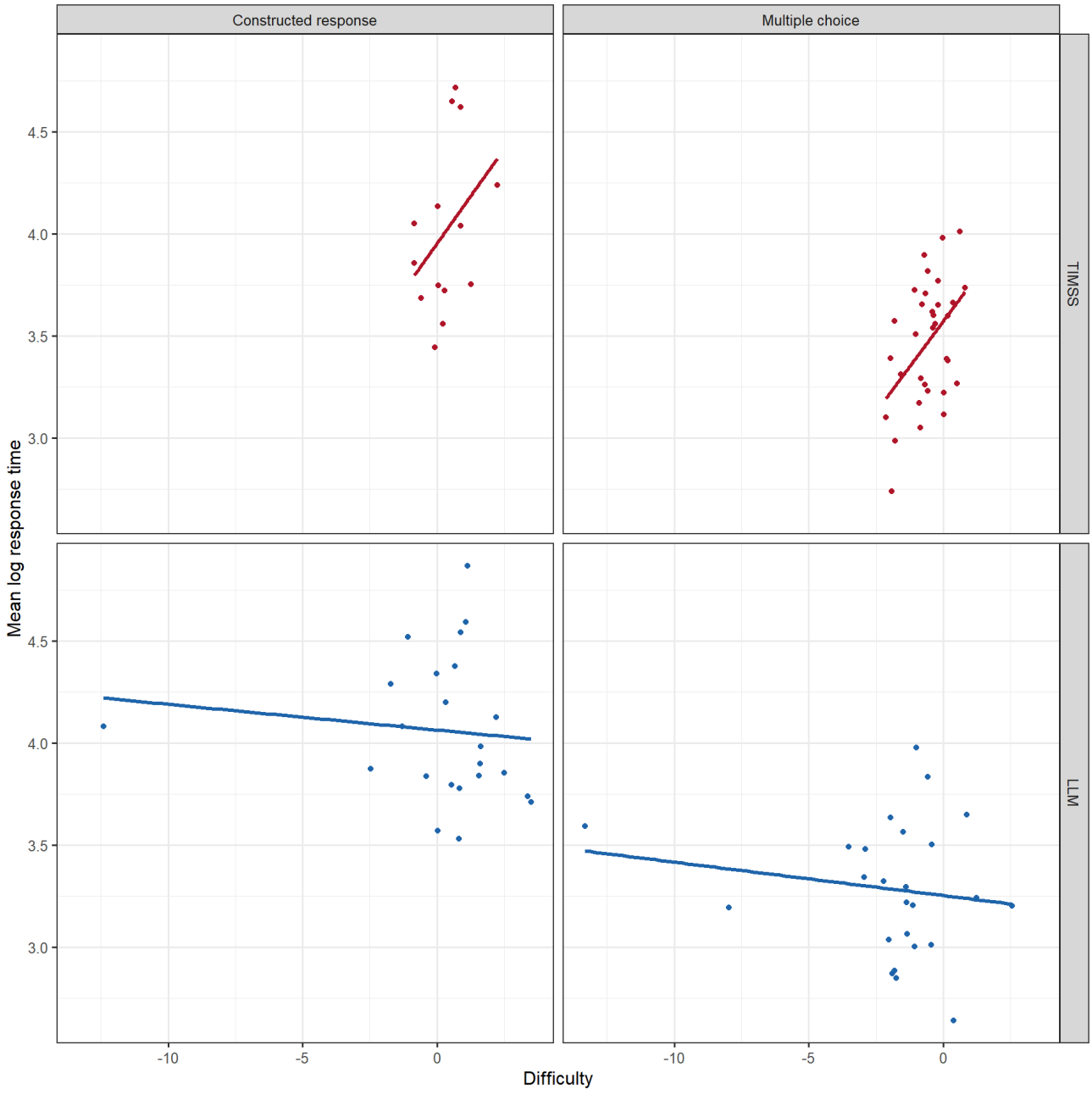
	Kruskal-Wallis rank sum test	ANOVA
Discrimination	$\chi^2(3) = 1.76$, $p = 0.624$	$F(3, 53) = 0.58$, $p = 0.629$
Difficulty	$\chi^2(3) = 0.89$, $p = 0.827$	$F(3, 53) = 2.62$, $p = 0.060$
Mean log resp. time	$\chi^2(3) = 8.30$, $p = 0.040$	$F(3, 53) = 3.54$, $p = 0.021$

Comparison	Wilcoxon rnak sum exact test	t-test
Discrimination		
TIMSS vs. LLM (all items)	W = 419.00, p = 0.768	t(50.6) = -0.02, p = 0.980
TIMSS vs. LLM simple setup	W = 116.00, p = 0.908	t(9.6) = -0.16, p = 0.878
TIMSS vs. LLM complex setup	W = 144.00, p = 0.231	t(15.9) = 1.49, p = 0.155
TIMSS vs. LLM teachers' selection	W = 159.00, p = 0.654	t(16.4) = -0.49, p = 0.627
LLM simple setup vs. LLM complex setup	W = 47.00, p = 0.605	t(11.2) = 0.90, p = 0.385
LLM simple setup vs. LLM teachers' selection	W = 56.00, p = 0.688	t(18.4) = -0.22, p = 0.826

Comparison	Wilcoxon rnak sum exact test	t-test
LLM complex setup vs. LLM teachers' selection	W = 45.00, p = 0.277	t(18.8) = -1.30, p = 0.208
Difficulty		
TIMSS vs. LLM (all items)	W = 400.00, p = 1.000	t(43.4) = -0.82, p = 0.418
TIMSS vs. LLM simple setup	W = 96.00, p = 0.539	t(8.5) = -1.49, p = 0.173
TIMSS vs. LLM complex setup	W = 105.00, p = 0.788	t(10.6) = 0.24, p = 0.815
TIMSS vs. LLM teachers' selection	W = 199.00, p = 0.496	t(18.3) = 0.38, p = 0.711
LLM simple setup vs. LLM complex setup	W = 45.00, p = 0.730	t(10.8) = 1.49, p = 0.165
LLM simple setup vs. LLM teachers' selection	W = 73.00, p = 0.557	t(10.3) = 1.56, p = 0.149
LLM complex setup vs. LLM teachers' selection	W = 71.00, p = 0.643	t(18.5) = 0.08, p = 0.937
Mean log resp. time		
TIMSS vs. LLM (all items)	W = 537.00, p = 0.027	t(50.2) = 2.55, p = 0.014
TIMSS vs. LLM simple setup	W = 124.00, p = 0.673	t(13.1) = 0.60, p = 0.556
TIMSS vs. LLM complex setup	W = 144.00, p = 0.231	t(16.7) = 1.52, p = 0.148
TIMSS vs. LLM teachers' selection	W = 269.00, p = 0.005	t(31.8) = 3.43, p = 0.002
LLM simple setup vs. LLM complex setup	W = 47.00, p = 0.605	t(15) = 0.59, p = 0.565
LLM simple setup vs. LLM teachers' selection	W = 91.00, p = 0.083	t(13.7) = 1.79, p = 0.095
LLM complex setup vs. LLM teachers' selection	W = 83.00, p = 0.224	t(16.8) = 1.38, p = 0.185

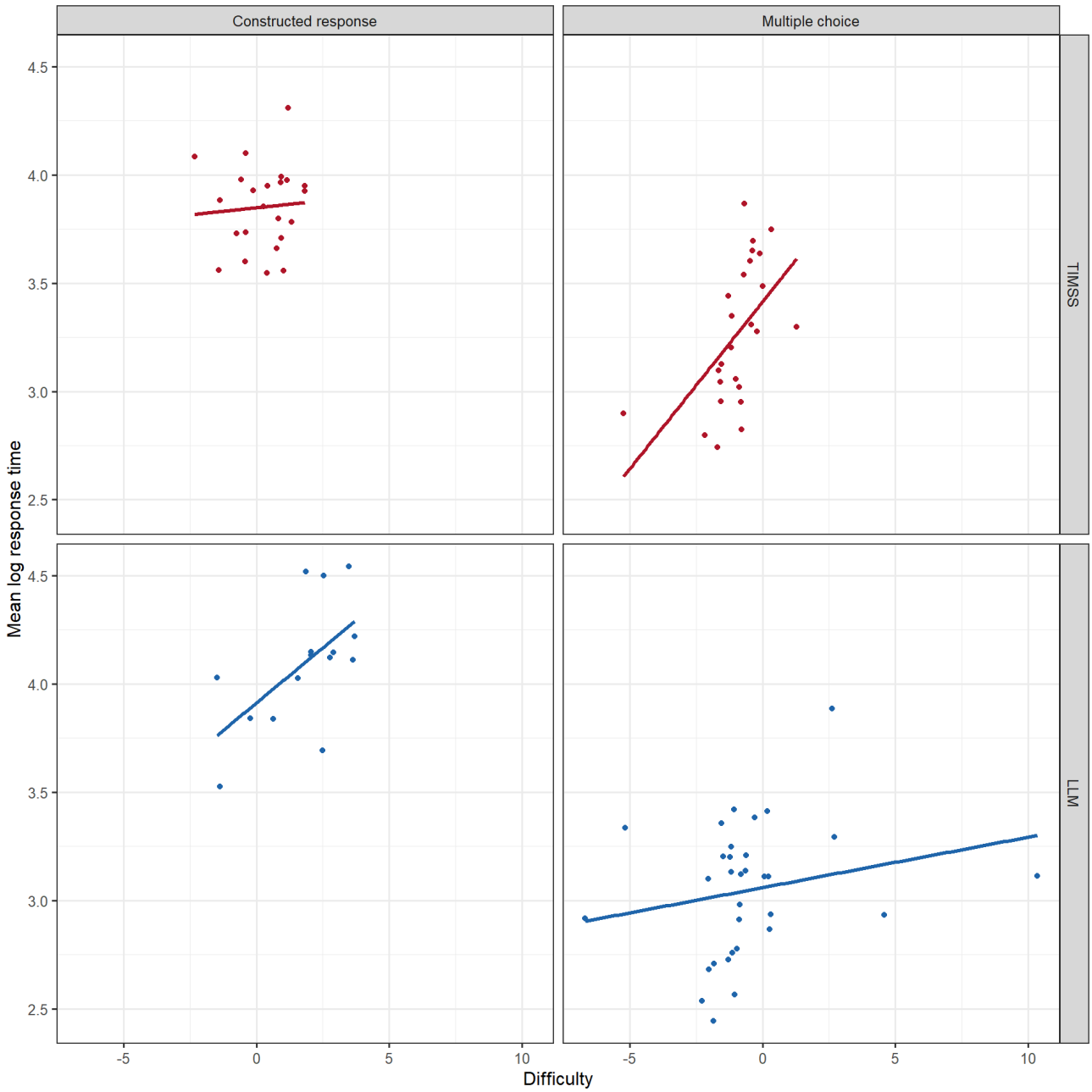
4 Correlations between item difficulty and response time

4.1 Math items



Group of items	Pearson correlation	t-test
Constructed response, TIMSS	0.38	t(12) = 1.42, p = 0.182
Multiple choice, TIMSS	0.46	t(31) = 2.88, p = 0.007
Constructed response, LLM	−0.11	t(21) = -0.51, p = 0.613
Multiple choice, LLM	−0.15	t(23) = -0.74, p = 0.467

4.2 Science items



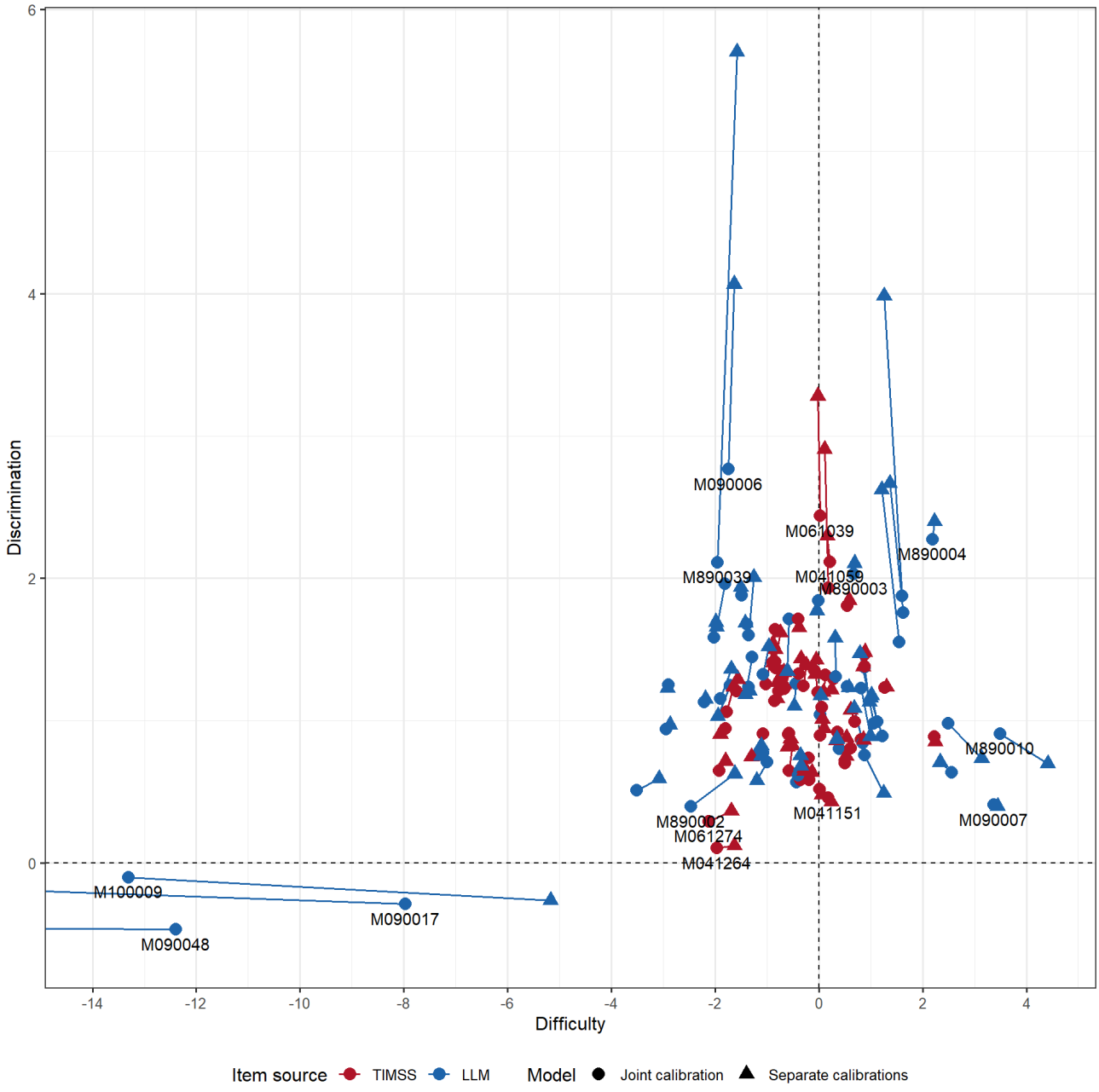
Group of items	Pearson correlation	t-test
Constructed response, TIMSS	0.07	t(21) = 0.33, p = 0.744
Multiple choice, TIMSS	0.55	t(23) = 3.14, p = 0.005
Constructed response, LLM	0.59	t(13) = 2.61, p = 0.022
Multiple choice, LLM	0.21	t(30) = 1.19, p = 0.244

5 Differences between models in parameter values

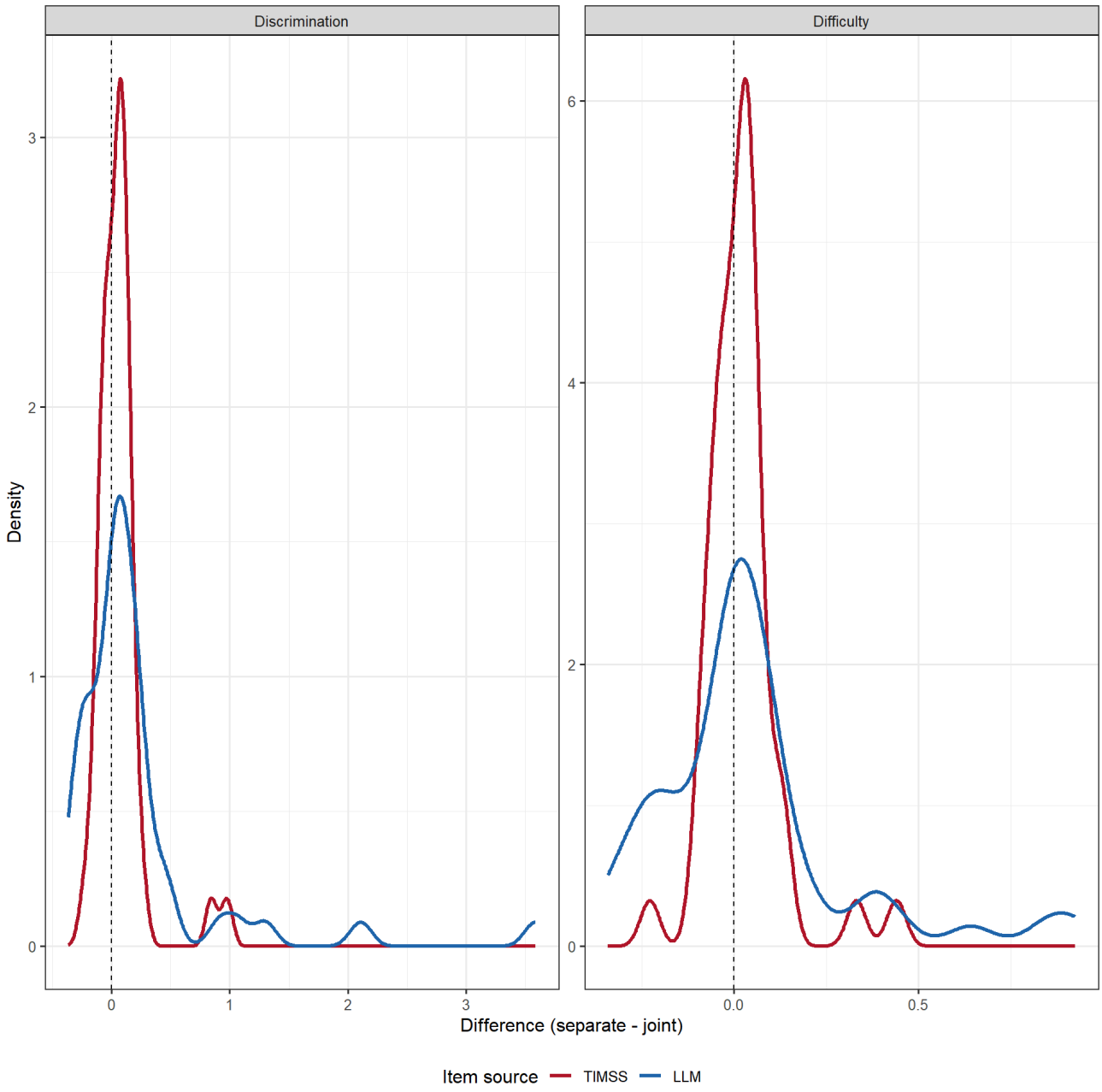
5.1 Unidimensional model for all items vs. separate unidimensional TIMSS model and LLM model

5.1.1 Math items

Source	Model	N	Min	Q1	Median	Q3	Max	Mean	SD
Discrimination									
LLM	Joint calibration	48	−0.47	0.79	1.19	1.59	2.77	1.16	0.64
LLM	Separate calibrations	48	−0.26	0.75	1.17	1.67	5.70	1.37	1.07
TIMSS	Joint calibration	47	0.10	0.88	1.20	1.35	2.44	1.12	0.46
TIMSS	Separate calibrations	47	0.12	0.84	1.22	1.42	3.28	1.19	0.58
Difficulty									
LLM	Joint calibration	48	−13.31	−1.76	−0.51	0.88	3.49	−0.92	3.22
LLM	Separate calibrations	48	−233.43	−1.62	−0.54	0.98	4.41	−5.62	33.79
TIMSS	Joint calibration	47	−2.13	−0.84	−0.30	0.20	2.23	−0.30	0.89
TIMSS	Separate calibrations	47	−1.90	−0.80	−0.27	0.20	2.25	−0.28	0.86



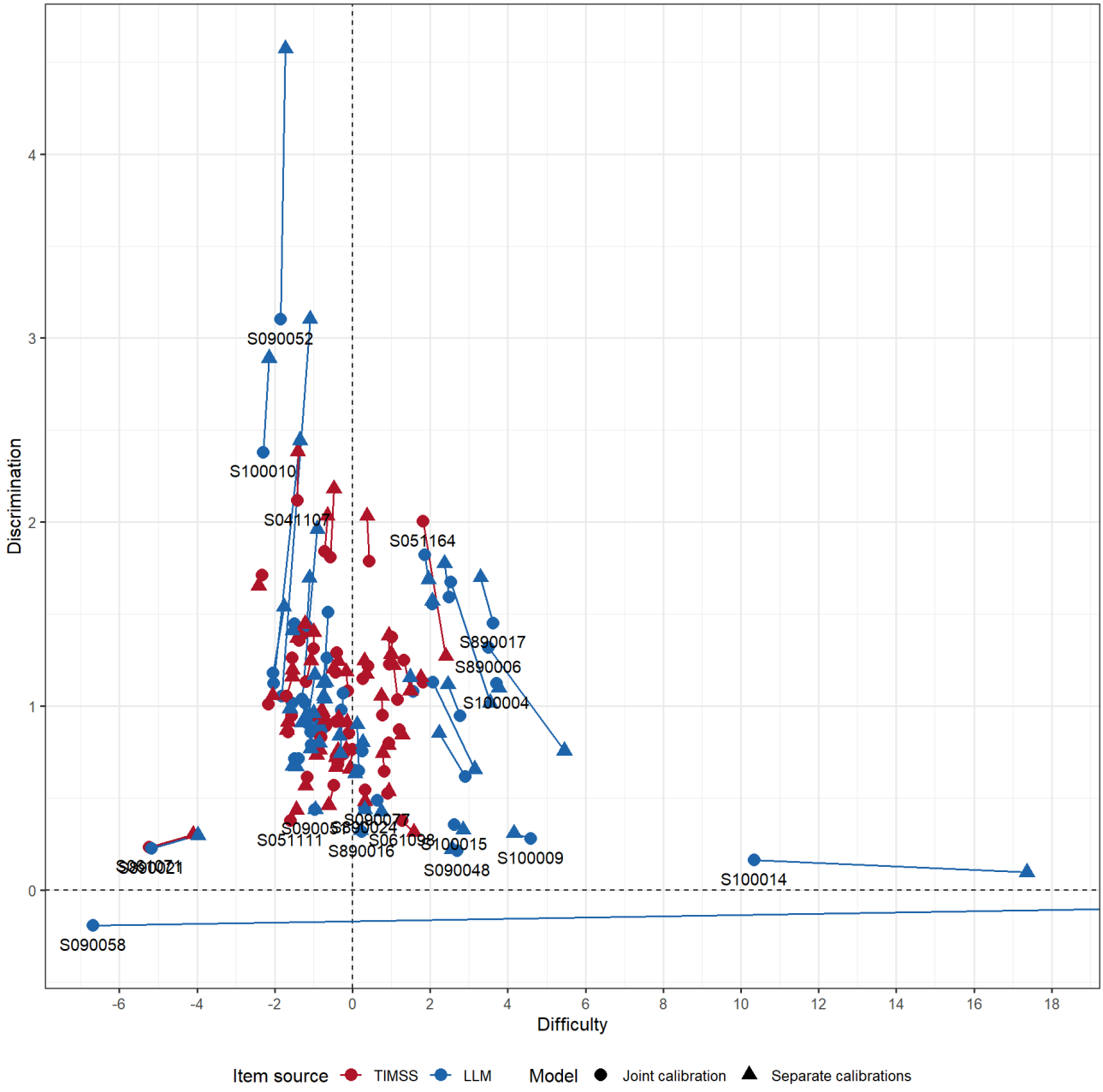

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_density()`).
```



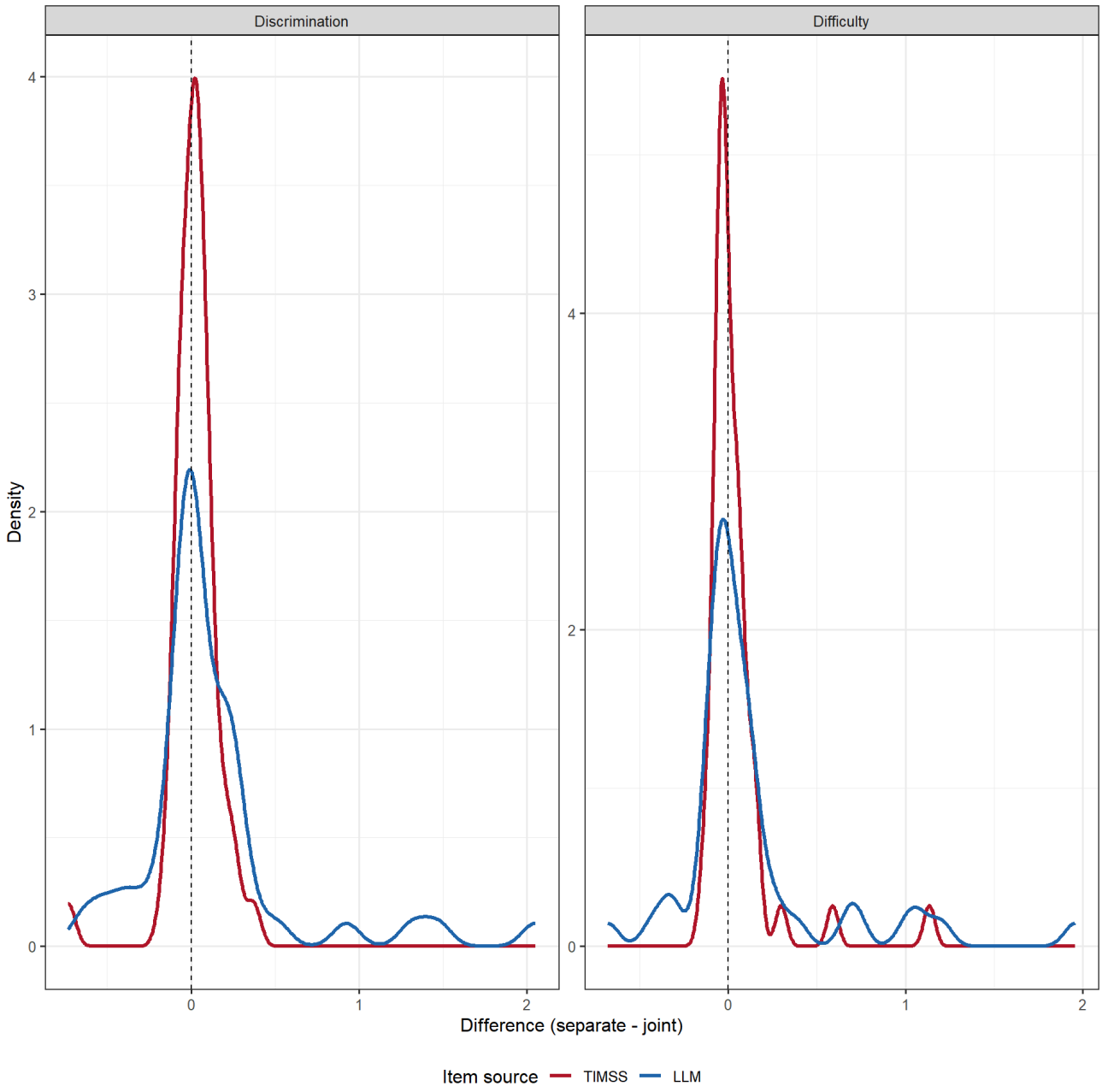
5.1.2 Science items

Source	Model	N	Min	Q1	Median	Q3	Max	Mean	SD
Discrimination									
LLM	Joint calibration	47	−0.19	0.65	0.95	1.22	3.10	0.98	0.58
LLM	Separate calibrations	47	0.02	0.66	0.92	1.29	4.57	1.11	0.84
TIMSS	Joint calibration	48	0.23	0.76	0.95	1.25	2.12	1.04	0.42
TIMSS	Separate calibrations	48	0.30	0.75	1.01	1.25	2.38	1.06	0.46
Difficulty									
LLM	Joint calibration	47	−6.68	−1.26	−0.63	2.06	10.35	0.21	2.71

Source	Model	N	Min	Q1	Median	Q3	Max	Mean	SD
LLM	Separate calibrations	47	−3.97	−1.10	−0.34	2.30	55.74	1.81	8.65
TIMSS	Joint calibration	48	−5.24	−1.17	−0.42	0.51	1.82	−0.38	1.27
TIMSS	Separate calibrations	48	−4.10	−1.10	−0.42	0.47	2.40	−0.34	1.22




```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```

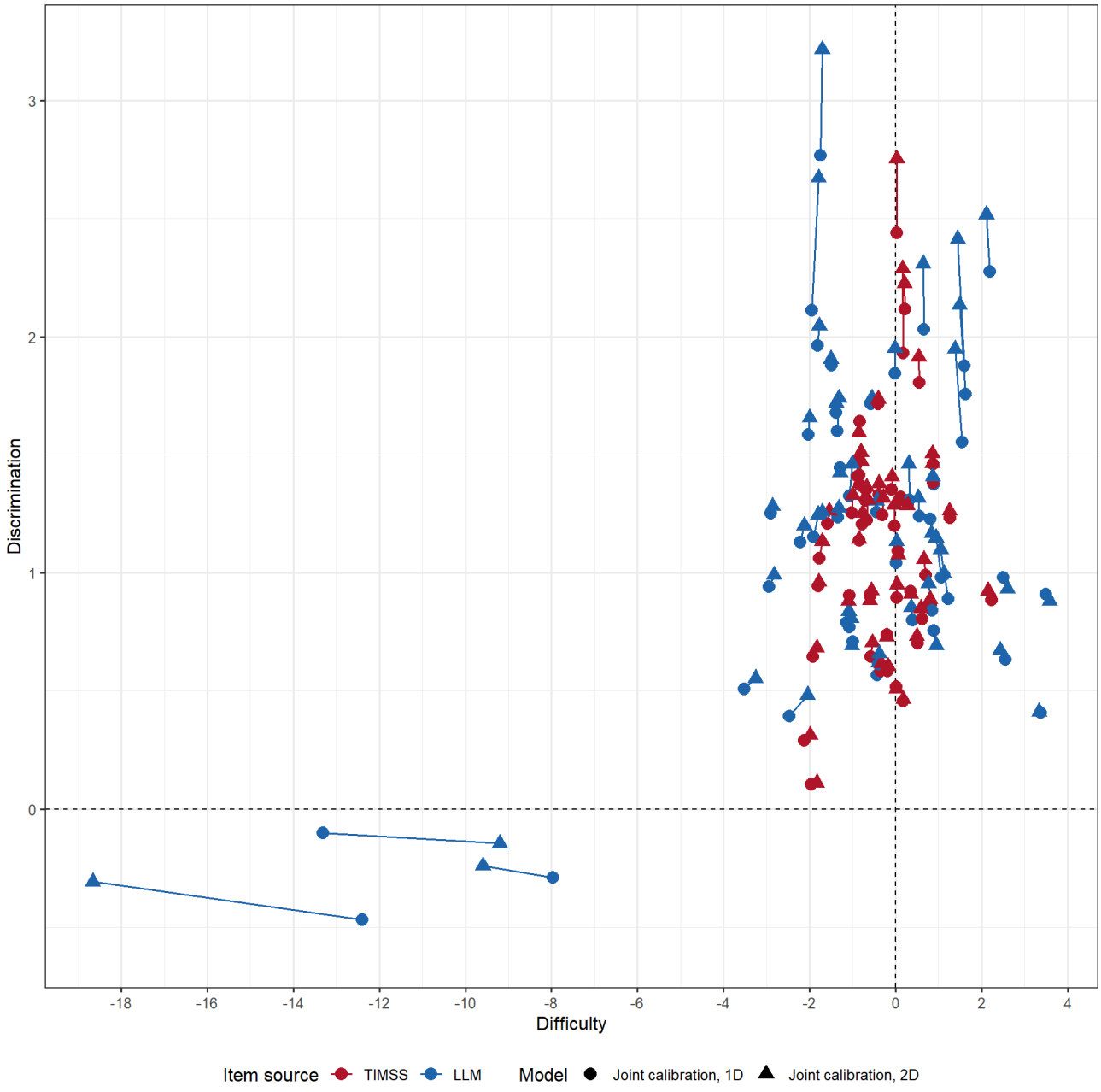


5.2 Unidimensional model for all items vs. twodimensional model for all items

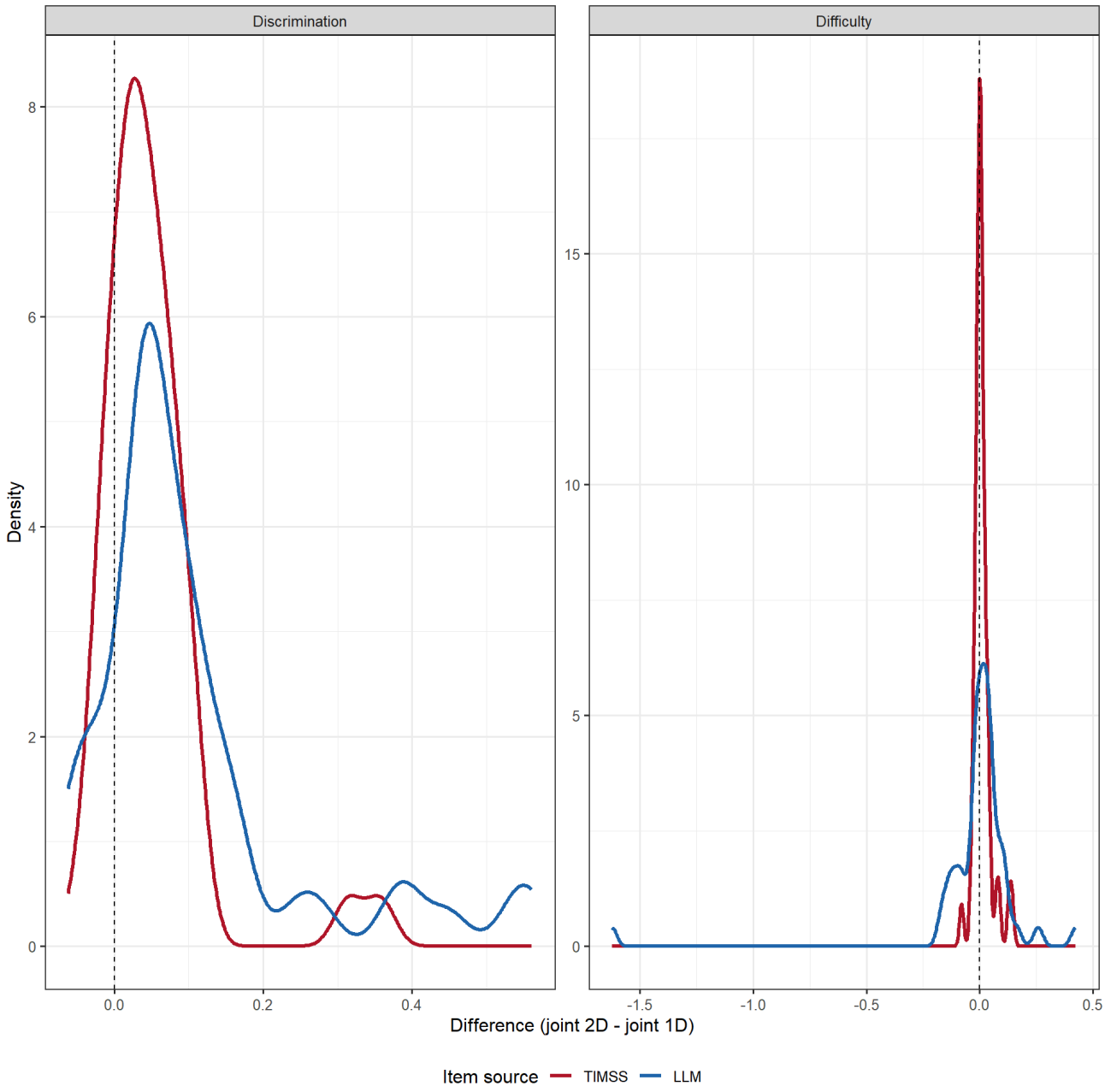
5.2.1 Math items

Source	Model	N	Min	Q1	Median	Q3	Max	Mean	SD
Discrimination									
LLM	Joint calibration, 1D	48	−0.47	0.79	1.19	1.59	2.77	1.16	0.64
LLM	Joint calibration, 2D	48	−0.31	0.83	1.22	1.72	3.22	1.26	0.73
TIMSS	Joint calibration, 1D	47	0.10	0.88	1.20	1.35	2.44	1.12	0.46
TIMSS	Joint calibration, 2D	47	0.11	0.88	1.25	1.39	2.75	1.17	0.50

Source	Model	N	Min	Q1	Median	Q3	Max	Mean	SD
Difficulty									
LLM	Joint calibration, 1D	48	−13.31	−1.76	−0.51	0.88	3.49	−0.92	3.22
LLM	Joint calibration, 2D	48	−18.67	−1.72	−0.49	0.89	3.59	−0.98	3.58
TIMSS	Joint calibration, 1D	47	−2.13	−0.84	−0.30	0.20	2.23	−0.30	0.89
TIMSS	Joint calibration, 2D	47	−1.98	−0.80	−0.29	0.20	2.15	−0.29	0.86



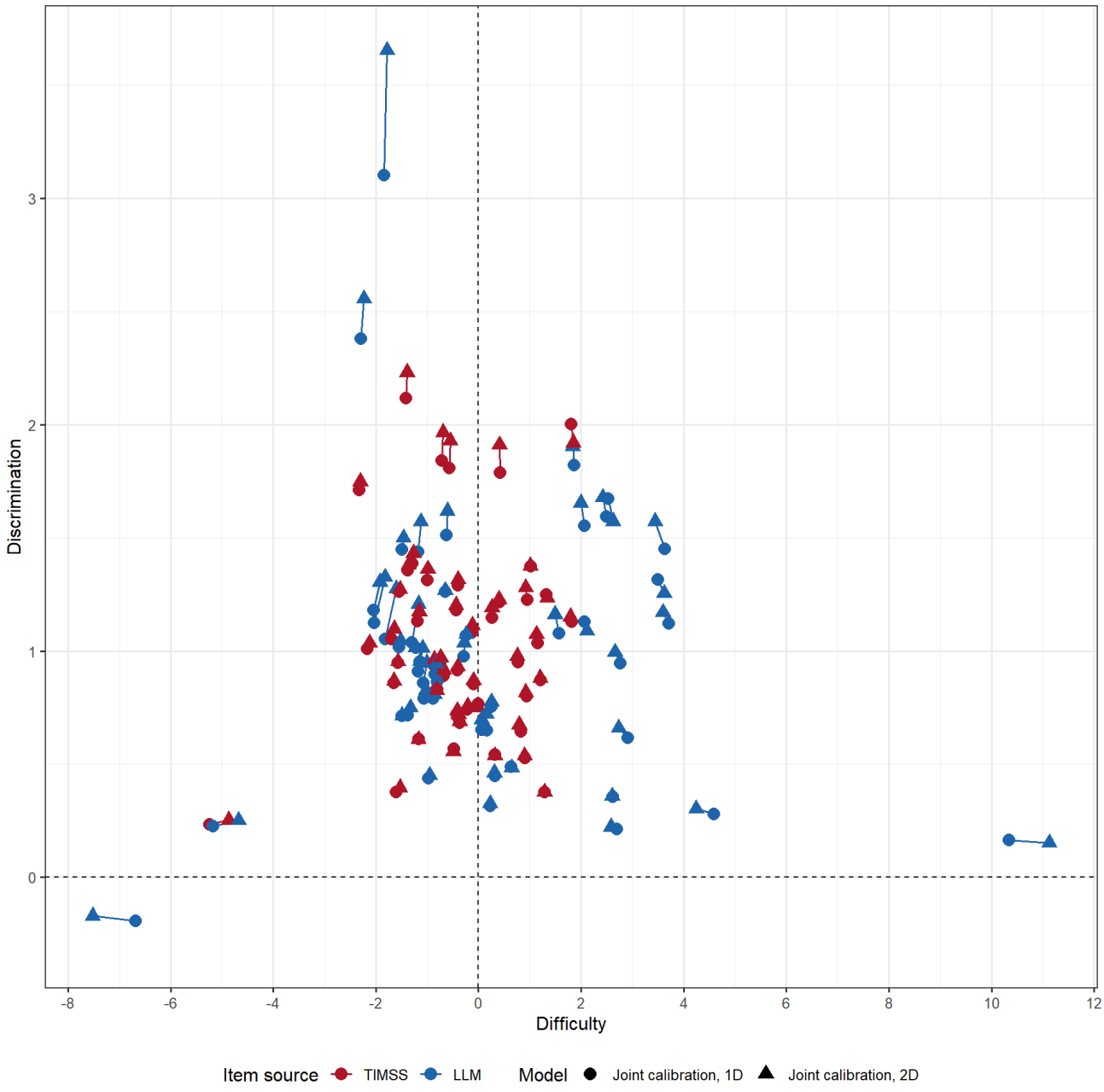

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```

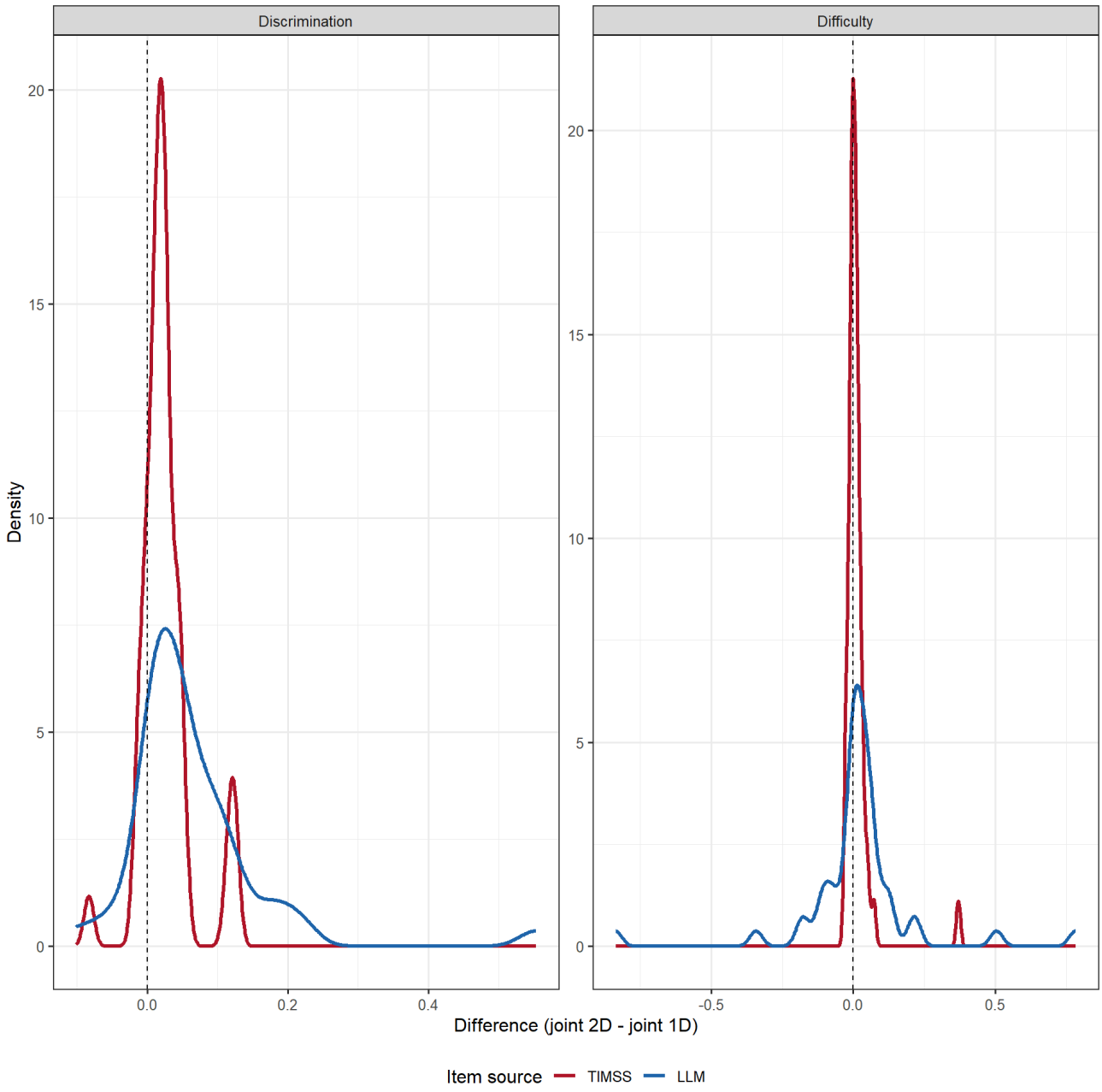


5.2.2 Science items

Source	Model	N	Min	Q1	Median	Q3	Max	Mean	SD
Discrimination									
LLM	Joint calibration, 1D	47	−0.19	0.65	0.95	1.22	3.10	0.98	0.58
LLM	Joint calibration, 2D	47	−0.17	0.71	1.01	1.29	3.65	1.05	0.65
TIMSS	Joint calibration, 1D	48	0.23	0.76	0.95	1.25	2.12	1.04	0.42
TIMSS	Joint calibration, 2D	48	0.25	0.76	0.97	1.27	2.23	1.06	0.44
Difficulty									
LLM	Joint calibration, 1D	47	−6.68	−1.26	−0.63	2.06	10.35	0.21	2.71

Source	Model	N	Min	Q1	Median	Q3	Max	Mean	SD
LLM	Joint calibration, 2D	47	−7.52	−1.19	−0.61	2.06	11.13	0.22	2.76
TIMSS	Joint calibration, 1D	48	−5.24	−1.17	−0.42	0.51	1.82	−0.38	1.27
TIMSS	Joint calibration, 2D	48	−4.87	−1.16	−0.41	0.50	1.85	−0.37	1.24


Discrimination
3
2
1
0
-8
-6
-4
-2
0
2
4
6
8
10
12
Difficulty
Item source
TIMSS
LLM
Model
Joint calibration, 1D
Joint calibration, 2D



6 Gender DIF

The logistic regression method (Swaminathan and Rogers, 1990) was used to detect both uniform and nonuniform gender DIF. Because of the missing by design property of the data the EAP ability estimates (from unidimensional models estimated separately for math and science, which included both TIMSS and LLM-generated item) where used as the matching criterion. Items with partial credit scoring rules were recoded into binary scores prior to the analysis by merging the middle score to the least frequently occuring of the remaining scores.

Because of the large number of performed tests Holm's correction for multiple comparisons was applied - separately to math items and science items DIF testing results. Jodoin and Gierl's (2001) criteria were used to categorise DIF effect size for the items for which DIF was statistically significant at the 0.05 significance level. Direction of DIF effect was reported on the basis of logistic regression coefficients. Analysis was performed using *difR* R package (Magis et al., 2010).

6.1 Math items

Source	Item	Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)	More difficult for	More discriminatir for
LLM	M090048	1.43	0.489	1.000	0.107	N/A	N/A	N/A
TIMSS	M041135	12.53	0.002	0.181	0.043	N/A	N/A	N/A
LLM	M890010	3.80	0.149	1.000	0.034	N/A	N/A	N/A
LLM	M890004	2.89	0.235	1.000	0.034	N/A	N/A	N/A
LLM	M890005	11.32	0.003	0.324	0.033	N/A	N/A	N/A
TIMSS	M061248	11.59	0.003	0.286	0.028	N/A	N/A	N/A
LLM	M090000	7.30	0.026	1.000	0.028	N/A	N/A	N/A
LLM	M100020	3.05	0.217	1.000	0.026	N/A	N/A	N/A
TIMSS	M051061	10.02	0.007	0.614	0.026	N/A	N/A	N/A
LLM	M090042	5.37	0.068	1.000	0.025	N/A	N/A	N/A
TIMSS	M041333	8.70	0.013	1.000	0.024	N/A	N/A	N/A
LLM	M090016	8.18	0.017	1.000	0.024	N/A	N/A	N/A
TIMSS	M041271	8.59	0.014	1.000	0.022	N/A	N/A	N/A
TIMSS	MP71217	8.56	0.014	1.000	0.022	N/A	N/A	N/A
TIMSS	M061167	7.06	0.029	1.000	0.019	N/A	N/A	N/A
LLM	M090007	5.06	0.080	1.000	0.018	N/A	N/A	N/A
TIMSS	M061106	6.21	0.045	1.000	0.018	N/A	N/A	N/A
LLM	M100010	5.12	0.077	1.000	0.014	N/A	N/A	N/A
TIMSS	M041059	5.81	0.055	1.000	0.014	N/A	N/A	N/A
LLM	M890039	1.97	0.373	1.000	0.014	N/A	N/A	N/A
TIMSS	M051205	4.50	0.105	1.000	0.013	N/A	N/A	N/A
LLM	M890029	2.77	0.250	1.000	0.013	N/A	N/A	N/A
TIMSS	M061236	3.51	0.173	1.000	0.012	N/A	N/A	N/A

Source	Item	Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)	More difficult for	More discriminatir for
LLM	M100028	2.25	0.324	1.000	0.011	N/A	N/A	N/A
LLM	M100009	2.98	0.225	1.000	0.011	N/A	N/A	N/A
TIMSS	M041048	3.82	0.148	1.000	0.011	N/A	N/A	N/A
TIMSS	M051236	2.56	0.278	1.000	0.011	N/A	N/A	N/A
LLM	M890003	3.49	0.175	1.000	0.010	N/A	N/A	N/A
TIMSS	M051039	3.64	0.162	1.000	0.010	N/A	N/A	N/A
LLM	M890024	1.69	0.429	1.000	0.010	N/A	N/A	N/A
TIMSS	M061207	3.73	0.155	1.000	0.010	N/A	N/A	N/A
TIMSS	M041177	2.29	0.318	1.000	0.009	N/A	N/A	N/A
LLM	M090047	2.87	0.238	1.000	0.008	N/A	N/A	N/A
TIMSS	M051077	3.02	0.221	1.000	0.008	N/A	N/A	N/A
TIMSS	M051070	2.71	0.258	1.000	0.008	N/A	N/A	N/A
LLM	M090057	2.68	0.261	1.000	0.008	N/A	N/A	N/A
LLM	M890011	1.91	0.384	1.000	0.007	N/A	N/A	N/A
TIMSS	M061039	2.63	0.268	1.000	0.006	N/A	N/A	N/A
TIMSS	M041268	2.27	0.321	1.000	0.006	N/A	N/A	N/A
TIMSS	M041174	2.16	0.339	1.000	0.006	N/A	N/A	N/A
LLM	M890006	2.02	0.364	1.000	0.006	N/A	N/A	N/A
TIMSS	M061052	1.95	0.377	1.000	0.006	N/A	N/A	N/A
TIMSS	M061050	1.95	0.377	1.000	0.006	N/A	N/A	N/A
LLM	M890025	1.83	0.400	1.000	0.005	N/A	N/A	N/A
LLM	M890021	1.16	0.560	1.000	0.005	N/A	N/A	N/A
LLM	M890030	1.56	0.459	1.000	0.005	N/A	N/A	N/A
TIMSS	M061274	1.49	0.475	1.000	0.004	N/A	N/A	N/A

Source	Item	Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)	More difficult for	More discriminatir for
LLM	M090013	1.53	0.466	1.000	0.004	N/A	N/A	N/A
TIMSS	M061179	1.42	0.491	1.000	0.004	N/A	N/A	N/A
TIMSS	M041169	1.50	0.473	1.000	0.004	N/A	N/A	N/A
LLM	M890002	1.20	0.550	1.000	0.004	N/A	N/A	N/A
TIMSS	M041298	0.97	0.616	1.000	0.004	N/A	N/A	N/A
LLM	M090017	0.62	0.733	1.000	0.004	N/A	N/A	N/A
LLM	M090012	0.58	0.750	1.000	0.004	N/A	N/A	N/A
LLM	M100017	0.81	0.666	1.000	0.004	N/A	N/A	N/A
TIMSS	M041151	1.25	0.536	1.000	0.003	N/A	N/A	N/A
LLM	M100008	0.96	0.618	1.000	0.003	N/A	N/A	N/A
LLM	M090059	0.85	0.655	1.000	0.003	N/A	N/A	N/A
TIMSS	M041034	1.16	0.560	1.000	0.003	N/A	N/A	N/A
TIMSS	M061206	1.10	0.576	1.000	0.003	N/A	N/A	N/A
TIMSS	M041280	1.03	0.596	1.000	0.003	N/A	N/A	N/A
LLM	M090044	0.69	0.708	1.000	0.003	N/A	N/A	N/A
TIMSS	M061272	0.99	0.610	1.000	0.003	N/A	N/A	N/A
LLM	M890026	0.86	0.651	1.000	0.002	N/A	N/A	N/A
TIMSS	M041191	0.57	0.754	1.000	0.002	N/A	N/A	N/A
LLM	M100011	0.65	0.724	1.000	0.002	N/A	N/A	N/A
LLM	M890009	0.78	0.678	1.000	0.002	N/A	N/A	N/A
LLM	M090055	0.34	0.845	1.000	0.002	N/A	N/A	N/A
LLM	M100027	0.49	0.783	1.000	0.002	N/A	N/A	N/A
TIMSS	MP71021	0.63	0.730	1.000	0.002	N/A	N/A	N/A
LLM	M100037	0.38	0.827	1.000	0.002	N/A	N/A	N/A

Source	Item	Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)	More difficult for	More discriminatir for
TIMSS	M051129	0.52	0.769	1.000	0.002	N/A	N/A	N/A
TIMSS	MP71219	0.47	0.792	1.000	0.001	N/A	N/A	N/A
TIMSS	MP71041	0.45	0.797	1.000	0.001	N/A	N/A	N/A
LLM	M100013	0.55	0.758	1.000	0.001	N/A	N/A	N/A
TIMSS	MP71162	0.44	0.801	1.000	0.001	N/A	N/A	N/A
LLM	M090051	0.48	0.785	1.000	0.001	N/A	N/A	N/A
LLM	M100026	0.44	0.802	1.000	0.001	N/A	N/A	N/A
LLM	M100039	0.44	0.802	1.000	0.001	N/A	N/A	N/A
TIMSS	MP71090	0.40	0.817	1.000	0.001	N/A	N/A	N/A
TIMSS	M041264	0.42	0.812	1.000	0.001	N/A	N/A	N/A
TIMSS	M061185	0.36	0.835	1.000	0.001	N/A	N/A	N/A
LLM	M090006	0.13	0.937	1.000	0.001	N/A	N/A	N/A
TIMSS	M041096	0.30	0.860	1.000	0.001	N/A	N/A	N/A
LLM	M890022	0.26	0.877	1.000	0.001	N/A	N/A	N/A
LLM	M100024	0.26	0.880	1.000	0.001	N/A	N/A	N/A
TIMSS	M061031	0.24	0.888	1.000	0.001	N/A	N/A	N/A
LLM	M100004	0.10	0.951	1.000	0.000	N/A	N/A	N/A
TIMSS	M051093	0.08	0.961	1.000	0.000	N/A	N/A	N/A
TIMSS	M041125	0.06	0.969	1.000	0.000	N/A	N/A	N/A
TIMSS	M041276	0.06	0.972	1.000	0.000	N/A	N/A	N/A
LLM	M090052	0.04	0.978	1.000	0.000	N/A	N/A	N/A
LLM	M100007	0.04	0.979	1.000	0.000	N/A	N/A	N/A
TIMSS	M061029	0.04	0.981	1.000	0.000	N/A	N/A	N/A
LLM	M100002	0.02	0.990	1.000	0.000	N/A	N/A	N/A

6.2 Science items

Source	Item	Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)	More difficult for	More discrimina for
LLM	S890009	19.69	0.000	0.005	0.058	moderate	males	males
LLM	S090051	17.99	0.000	0.012	0.051	moderate	females	males
TIMSS	S051121	17.64	0.000	0.014	0.051	moderate	females	males
TIMSS	S061031	17.06	0.000	0.018	0.049	moderate	females	males
LLM	S100032	12.31	0.002	0.193	0.105	N/A	N/A	N/A
LLM	S890017	9.81	0.007	0.637	0.084	N/A	N/A	N/A
TIMSS	S041033	11.88	0.003	0.237	0.049	N/A	N/A	N/A
LLM	S890038	11.19	0.004	0.330	0.048	N/A	N/A	N/A
TIMSS	S041010	10.43	0.005	0.477	0.045	N/A	N/A	N/A
LLM	S100010	3.11	0.211	1.000	0.036	N/A	N/A	N/A
TIMSS	S041183	7.93	0.019	1.000	0.032	N/A	N/A	N/A
LLM	S090052	3.76	0.153	1.000	0.031	N/A	N/A	N/A
LLM	S090018	9.45	0.009	0.754	0.029	N/A	N/A	N/A
TIMSS	S061138	10.12	0.006	0.552	0.028	N/A	N/A	N/A
TIMSS	S061105	8.85	0.012	0.996	0.026	N/A	N/A	N/A
LLM	S090080	7.15	0.028	1.000	0.026	N/A	N/A	N/A
LLM	S100014	6.18	0.045	1.000	0.025	N/A	N/A	N/A
TIMSS	S041081	8.94	0.011	0.964	0.025	N/A	N/A	N/A
LLM	S090041	6.67	0.036	1.000	0.024	N/A	N/A	N/A
LLM	S090077	8.48	0.014	1.000	0.024	N/A	N/A	N/A
LLM	S100024	5.70	0.058	1.000	0.023	N/A	N/A	N/A
TIMSS	S041027	2.67	0.263	1.000	0.023	N/A	N/A	N/A
TIMSS	S051105	6.88	0.032	1.000	0.022	N/A	N/A	N/A

Source	Item	Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)	More difficult for	More discrimina for
TIMSS	SP61011	7.47	0.024	1.000	0.021	N/A	N/A	N/A
TIMSS	SP71147	7.14	0.028	1.000	0.021	N/A	N/A	N/A
LLM	S890022	4.27	0.118	1.000	0.021	N/A	N/A	N/A
TIMSS	SP71268	6.83	0.033	1.000	0.019	N/A	N/A	N/A
TIMSS	S061071	4.90	0.086	1.000	0.017	N/A	N/A	N/A
LLM	S100004	1.14	0.566	1.000	0.016	N/A	N/A	N/A
TIMSS	S061034	4.87	0.088	1.000	0.015	N/A	N/A	N/A
TIMSS	SP71072	5.03	0.081	1.000	0.014	N/A	N/A	N/A
TIMSS	S051147	4.29	0.117	1.000	0.014	N/A	N/A	N/A
LLM	S100015	4.37	0.112	1.000	0.014	N/A	N/A	N/A
TIMSS	S061065	4.80	0.091	1.000	0.014	N/A	N/A	N/A
TIMSS	S051087	3.54	0.170	1.000	0.013	N/A	N/A	N/A
TIMSS	S041052	4.26	0.119	1.000	0.012	N/A	N/A	N/A
LLM	S890021	3.53	0.171	1.000	0.012	N/A	N/A	N/A
TIMSS	S041068	3.49	0.174	1.000	0.010	N/A	N/A	N/A
LLM	S890016	3.61	0.164	1.000	0.010	N/A	N/A	N/A
LLM	S890006	1.75	0.417	1.000	0.010	N/A	N/A	N/A
TIMSS	S051110	3.31	0.191	1.000	0.010	N/A	N/A	N/A
TIMSS	S041070	3.42	0.181	1.000	0.009	N/A	N/A	N/A
TIMSS	S061028	2.43	0.297	1.000	0.009	N/A	N/A	N/A
TIMSS	SP71017	2.75	0.253	1.000	0.008	N/A	N/A	N/A
TIMSS	S051188	2.69	0.261	1.000	0.008	N/A	N/A	N/A
TIMSS	S041032	1.94	0.380	1.000	0.008	N/A	N/A	N/A
TIMSS	S051037	2.54	0.281	1.000	0.007	N/A	N/A	N/A
LLM	S100008	2.00	0.369	1.000	0.007	N/A	N/A	N/A

Source	Item	Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)	More difficult for	More discrimina for
TIMSS	S041006	2.49	0.288	1.000	0.007	N/A	N/A	N/A
LLM	S890023	2.37	0.306	1.000	0.007	N/A	N/A	N/A
TIMSS	S041043	2.17	0.337	1.000	0.006	N/A	N/A	N/A
TIMSS	SP71054	1.93	0.382	1.000	0.006	N/A	N/A	N/A
LLM	S090048	1.96	0.376	1.000	0.006	N/A	N/A	N/A
TIMSS	SP71115	1.85	0.397	1.000	0.005	N/A	N/A	N/A
LLM	S090058	1.33	0.515	1.000	0.005	N/A	N/A	N/A
LLM	S090040	1.04	0.596	1.000	0.005	N/A	N/A	N/A
LLM	S090020	1.72	0.422	1.000	0.005	N/A	N/A	N/A
TIMSS	SP71402	1.37	0.504	1.000	0.005	N/A	N/A	N/A
LLM	S090064	1.47	0.480	1.000	0.005	N/A	N/A	N/A
TIMSS	S041102	1.29	0.525	1.000	0.005	N/A	N/A	N/A
LLM	S100029	1.34	0.512	1.000	0.004	N/A	N/A	N/A
TIMSS	S061060	1.42	0.492	1.000	0.004	N/A	N/A	N/A
LLM	S890002	0.68	0.712	1.000	0.004	N/A	N/A	N/A
LLM	S890005	0.42	0.810	1.000	0.004	N/A	N/A	N/A
LLM	S100016	1.25	0.534	1.000	0.004	N/A	N/A	N/A
TIMSS	S051111	1.22	0.544	1.000	0.004	N/A	N/A	N/A
TIMSS	SP71140	1.11	0.575	1.000	0.004	N/A	N/A	N/A
TIMSS	S061098	1.17	0.557	1.000	0.003	N/A	N/A	N/A
LLM	S890011	1.25	0.536	1.000	0.003	N/A	N/A	N/A
TIMSS	S041107	0.77	0.680	1.000	0.003	N/A	N/A	N/A
TIMSS	SP61044	0.97	0.614	1.000	0.003	N/A	N/A	N/A
TIMSS	SP71128	0.68	0.710	1.000	0.002	N/A	N/A	N/A
TIMSS	S061130	0.64	0.726	1.000	0.002	N/A	N/A	N/A

Source	Item	Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)	More difficult for	More discrimina for
LLM	S100009	0.63	0.730	1.000	0.002	N/A	N/A	N/A
TIMSS	S061010	0.80	0.669	1.000	0.002	N/A	N/A	N/A
LLM	S100012	0.52	0.770	1.000	0.002	N/A	N/A	N/A
LLM	S100002	0.53	0.766	1.000	0.002	N/A	N/A	N/A
TIMSS	S051164	0.24	0.887	1.000	0.002	N/A	N/A	N/A
LLM	S890007	0.30	0.863	1.000	0.001	N/A	N/A	N/A
LLM	S890010	0.48	0.787	1.000	0.001	N/A	N/A	N/A
LLM	S100039	0.30	0.859	1.000	0.001	N/A	N/A	N/A
LLM	S890026	0.39	0.824	1.000	0.001	N/A	N/A	N/A
TIMSS	S051114	0.40	0.820	1.000	0.001	N/A	N/A	N/A
LLM	S890004	0.31	0.858	1.000	0.001	N/A	N/A	N/A
LLM	S090084	0.34	0.844	1.000	0.001	N/A	N/A	N/A
TIMSS	S041209	0.21	0.898	1.000	0.001	N/A	N/A	N/A
LLM	S100023	0.11	0.949	1.000	0.001	N/A	N/A	N/A
LLM	S090075	0.15	0.926	1.000	0.001	N/A	N/A	N/A
LLM	S090015	0.12	0.941	1.000	0.000	N/A	N/A	N/A
TIMSS	SP71077	0.08	0.962	1.000	0.000	N/A	N/A	N/A
LLM	S090053	0.07	0.966	1.000	0.000	N/A	N/A	N/A
LLM	S090001	0.04	0.978	1.000	0.000	N/A	N/A	N/A
TIMSS	S051201	0.04	0.981	1.000	0.000	N/A	N/A	N/A
LLM	S890024	0.04	0.981	1.000	0.000	N/A	N/A	N/A
LLM	S100020	0.02	0.992	1.000	0.000	N/A	N/A	N/A

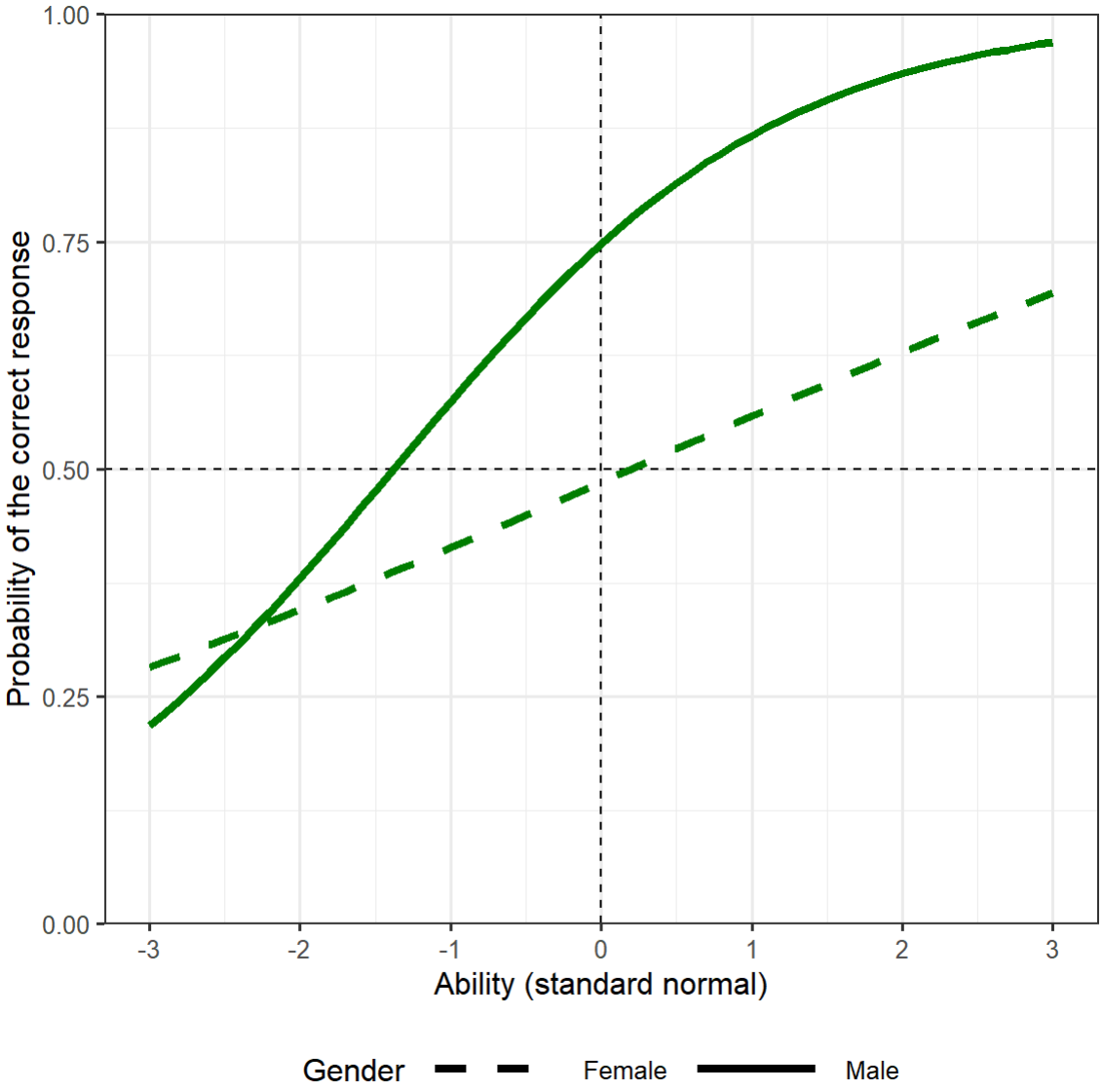
To illustarte consequences of the detected DIF a two-group unidimensional IRT 2PL model was estimated with:

1. Groups defined by gender

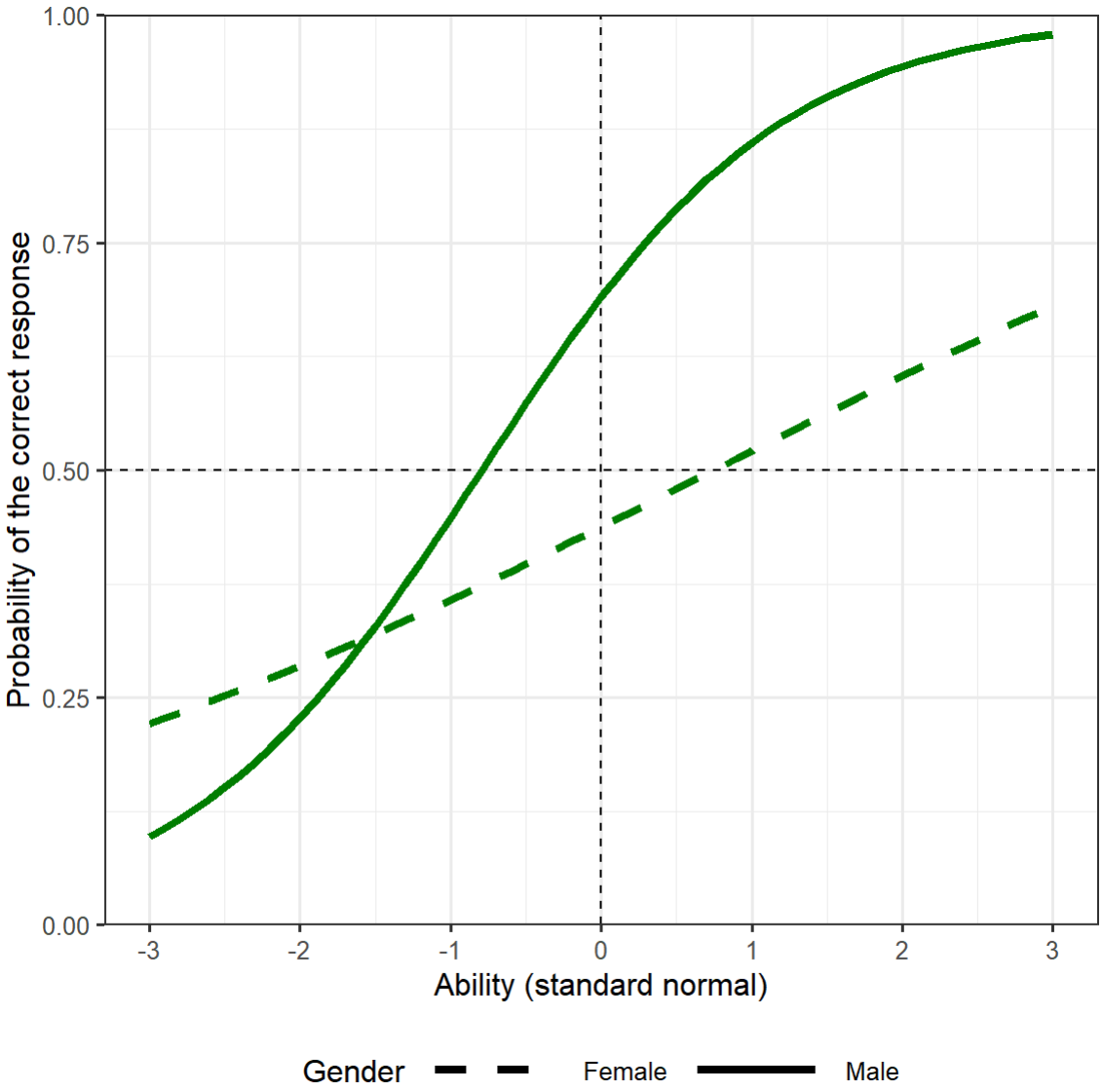
2. Latent trait distribution assumed to have expected value of 0 and variance of 1 for men but freely estimated expected value and variance for women
3. Both difficulty and discrimination parameters of items with detected DIF allowed to vary between groups
4. Values of all the other item held constant across groups

Below, ICCs and item parameters for items with detected DIF from this model are shown.

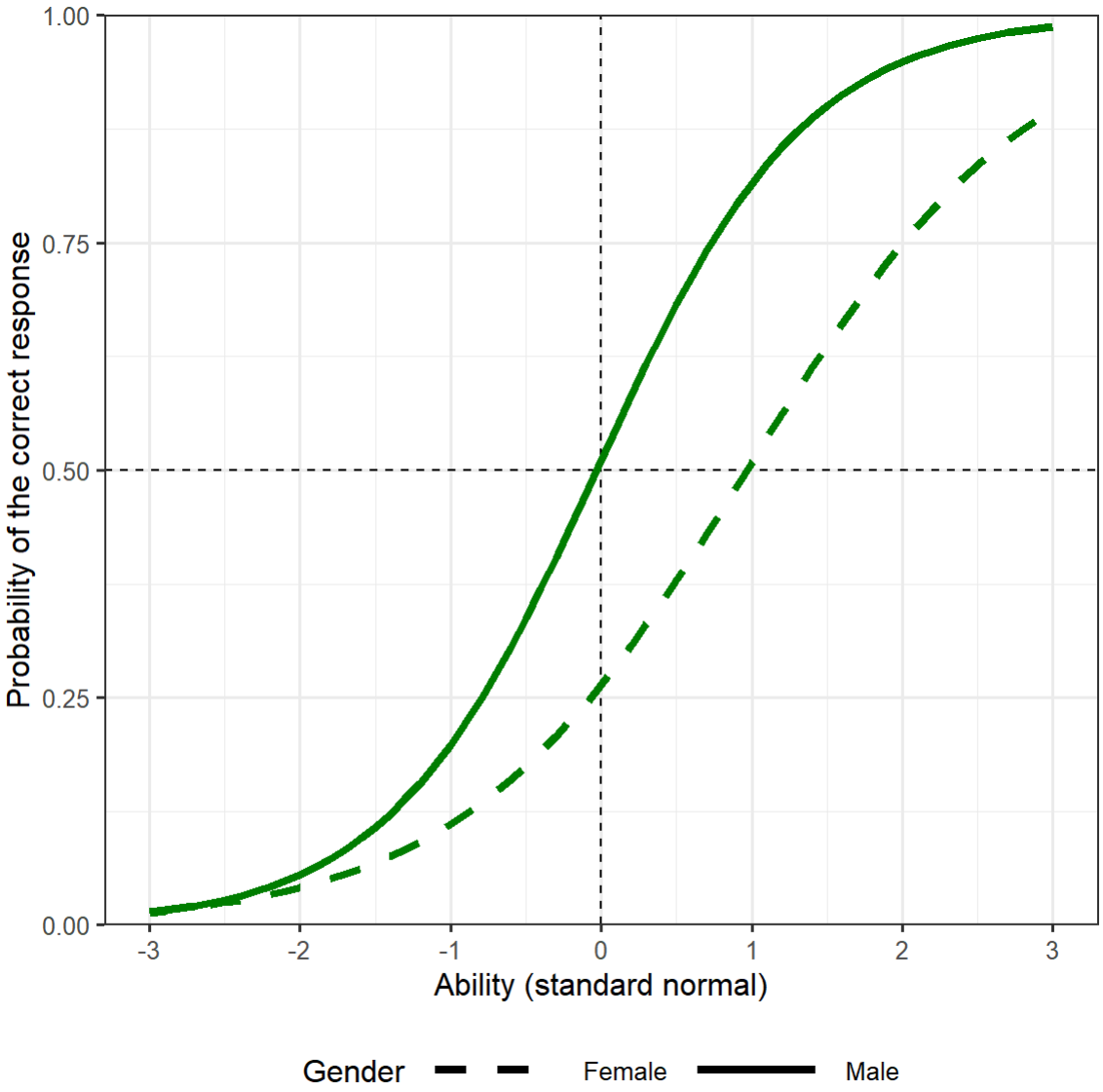
6.2.1 S090051



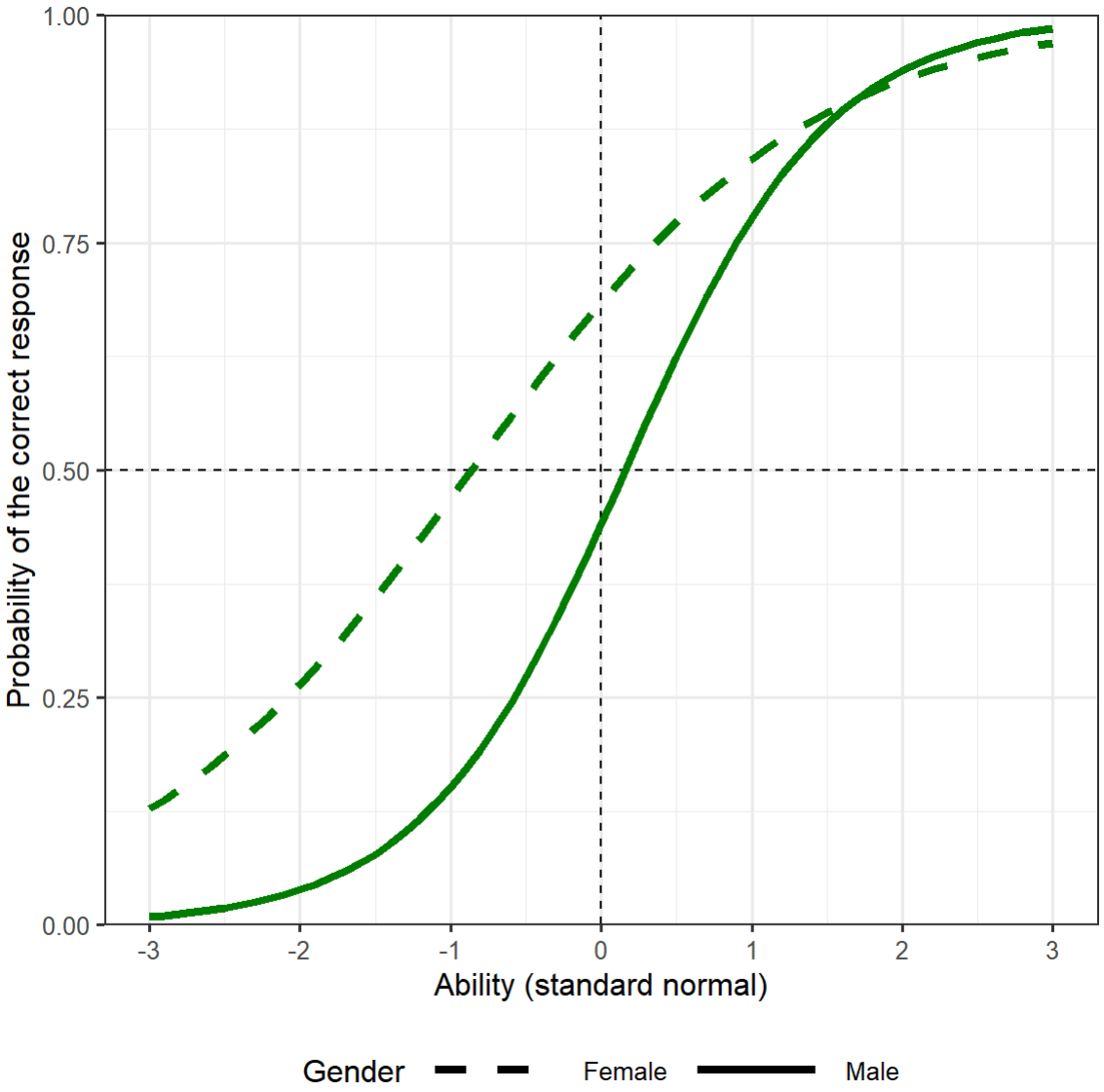
6.2.2 S051121


1.00
0.75
0.50
0.25
0.00
Probability of the correct response
-3
-2
-1
0
1
2
3
Ability (standard normal)
Gender
Female
Male

6.2.3 S061031


1.00
0.75
0.50
0.25
0.00
Probability of the correct response
-3
-2
-1
0
1
2
3
Ability (standard normal)
Gender
Female
Male

6.2.4 S890009



7 Local dependency diagnostics

Local dependencies were assed using the \(G^2\) statistic (Chen & Thissen, 1997) on the basis of the same unidimensional math and science 2PL IRT models involving both TIMSS and LLM-generated items, which parameters were described above. Please note that because of the missing by design property of the data, local dependency statistics can be computed only for the pairs of items that were placed in the same booklet (i.e. there are students who responded both items in a given pair). Consequently, local dependency statistics were computed for 1633 out of the total 4465 pairs of the calibrated math items and for 1645 out of the total 4465 pairs of the calibrated science items. Computations were performed using the *residuals()* function from the *mirt* R package (Chalmers, 2012).

Holm's correction for multiple comparisons was applied separately to the results of math item pairs local dependency tests and to the results of science item pairs local dependency tests. After the correction local dependencies signifficant at the 0.05 level are:

Item 1	Item 2	Correlation between standardised residuals	G^2	Df	Holm's adjusted p-value
Math					

Item 1	Item 2	Correlation between standardised residuals	G^2	Df	Holm's adjusted p-value
M100027 (LLM simple setup, CR)	M890030 (LLM complex setup, MC)	0.289	26.64	1	0.000
M890009 (LLM complex setup, CR)	M100024 (LLM simple setup, CR)	0.262	21.92	2	0.028
Science					
S090080 (LLM teachers' selection, MC)	S890038 (LLM complex setup, MC)	0.309	26.69	1	0.000
S890022 (LLM complex setup, MC)	S890038 (LLM complex setup, MC)	0.290	23.49	1	0.002

8 Items summaries

8.1 Math items

					$r_{score,\theta}$			IRT model includi	
Item	Format	Max. points	Number of obs.	Easiness	All items	TIMSS items only	Average response time [s]	Discrimination	diffi
TIMSS									
M041034	MC	1	306	48.52%	0.38	0.44	28.3	0.90	
M041048	MC	1	299	59.06%	0.25	0.39	56.7	0.65	–
M041059	CR	1	299	44.97%	0.60	0.62	42.7	2.12	
M041096	MC	1	306	41.31%	0.33	0.26	31.0	0.70	
M041125	MC	1	306	34.43%	0.34	0.37	55.3	0.86	
M041135	MC	1	299	76.17%	0.25	0.27	18.4	0.65	–
M041151	MC	1	306	47.54%	0.22	0.15	36.6	0.46	
M041169	MC	1	306	60.00%	0.35	0.38	29.8	0.91	–

					$r_{score,\theta}$			IRT model includi	
Item	Format	Max. points	Number of obs.	Easiness	All items	TIMSS items only	Average response time [s]	Discrimination	diffi
M041174	CR	1	304	71.38%	0.49	0.43	57.8	1.64	–
M041177	MC	1	299	81.54%	0.35	0.30	40.3	0.94	–
M041191	MC	1	294	82.65%	0.40	0.41	32.7	1.21	–
M041264	MC	1	304	54.93%	0.06	0.05	35.6	0.10	–
M041268	MC	1	299	43.96%	0.40	0.38	53.3	0.92	
M041271	MC	1	304	60.53%	0.54	0.56	41.8	1.72	–
M041276	CR	1	294	37.07%	0.38	0.46	132.6	0.99	
M041280	MC	1	304	49.34%	0.26	0.24	34.3	0.52	
M041298	MC	1	299	83.22%	0.34	0.39	23.6	1.06	–
M041333	MC	1	304	51.97%	0.27	0.27	53.2	0.58	–
M051039	CR	1	306	69.51%	0.48	0.48	70.5	1.41	–
M051061	CR	1	304	48.03%	0.39	0.42	48.3	1.09	
M051070	MC	1	294	40.14%	0.37	0.33	65.3	0.80	
M051077	MC	1	306	56.07%	0.44	0.47	40.9	1.25	–
M051093	MC	1	294	54.08%	0.32	0.28	51.1	0.74	–
M051129	MC	1	294	70.41%	0.39	0.31	47.8	0.91	–
M051205	CR	1	306	27.87%	0.47	0.50	72.1	1.38	
M051236	CR	1	294	15.65%	0.27	0.27	84.0	0.89	
M061029	MC	1	304	58.88%	0.52	0.49	42.5	1.33	–
M061031	MC	1	294	46.26%	0.59	0.68	45.1	1.93	
M061039	CR	1	299	50.67%	0.67	0.72	84.3	2.44	
M061050	MC	1	304	67.76%	0.46	0.41	24.8	1.14	–
M061052	MC	1	294	66.33%	0.45	0.45	47.7	1.22	–
M061106	MC	1	299	67.45%	0.50	0.48	57.0	1.31	–

					$r_{score,\theta}$			IRT model includi	
Item	Format	Max. points	Number of obs.	Easiness	All items	TIMSS items only	Average response time [s]	Discrimination	diffi
M061167	CR	1	304	60.53%	0.38	0.35	49.1	0.90	–
M061179	MC	1	304	72.37%	0.48	0.49	39.8	1.25	–
M061185	MC	1	306	65.57%	0.49	0.48	31.6	1.35	–
M061206	MC	1	306	54.10%	0.30	0.30	43.4	0.58	–
M061207	MC	1	299	47.99%	0.47	0.50	37.2	1.32	
M061236	CR	1	304	22.04%	0.40	0.41	51.5	1.23	
M061248	CR	2	306	34.26%	0.64	0.62	132.4	1.81	
M061272	CR	1	299	44.30%	0.50	0.46	51.4	1.29	
M061274	MC	1	304	64.47%	0.18	0.18	32.6	0.29	–
MP71021	MC	1	294	51.70%	0.46	0.54	67.0	1.20	–
MP71041	MC	1	294	71.09%	0.51	0.50	31.0	1.37	–
MP71090	MC	1	299	68.46%	0.45	0.48	47.4	1.21	–
MP71162	CR	1	294	29.59%	0.48	0.48	130.6	1.46	
MP71217	CR	1	306	51.15%	0.50	0.51	38.1	1.35	–
MP71219	MC	1	299	72.48%	0.49	0.51	27.9	1.41	–
LLM simple setup									
M100002	CR	2	319	31.35%	0.56	0.41	125.2	1.37	
M100004	MC	1	299	85.23%	0.36	0.21	20.9	1.15	–
M100007	CR	1	304	55.92%	0.30	0.18	57.9	0.61	–
M100008	MC	1	299	78.19%	0.43	0.27	29.0	1.23	–
M100009	MC	1	304	20.72%	−0.02	−0.01	44.4	−0.10	−1
M100010	MC	1	299	55.37%	0.24	0.15	44.0	0.57	–
M100011	CR	1	299	74.16%	0.47	0.25	111.2	1.33	–
M100013	CR	1	304	51.32%	0.55	0.36	97.2	1.84	–

					$r_{score,\theta}$			IRT model includi	
Item	Format	Max. points	Number of obs.	Easiness	All items	TIMSS items only	Average response time [s]	Discrimination	diffi
M100017	CR	1	299	12.42%	0.45	0.25	72.2	1.76	
M100020	MC	1	299	94.97%	0.27	0.20	36.3	1.25	–
M100024	CR	1	319	35.42%	0.37	0.20	57.0	0.84	
M100026	MC	1	304	35.86%	0.29	0.25	50.7	0.76	
M100027	CR	1	319	15.67%	0.47	0.25	62.1	1.55	
M100028	CR	1	304	10.86%	0.24	0.16	59.3	0.98	
M100037	MC	1	304	86.84%	0.49	0.31	39.2	1.88	–
M100039	MC	1	319	65.52%	0.29	0.26	66.2	0.71	–
LLM complex setup									
M890002	CR	1	281	71.79%	0.19	0.08	64.0	0.39	–
M890003	CR	1	281	30.00%	0.56	0.36	111.7	2.03	
M890004	CR	1	281	3.57%	0.31	0.17	94.9	2.27	
M890005	CR	1	281	36.79%	0.45	0.32	60.5	1.24	
M890006	CR	1	281	41.79%	0.47	0.30	97.4	1.31	
M890009	CR	2	319	33.23%	0.45	0.26	123.9	0.98	
M890010	CR	1	281	5.36%	0.15	0.15	56.0	0.91	
M890011	CR	1	319	84.33%	0.38	0.35	84.6	1.25	–
M890021	MC	1	319	84.64%	0.23	0.10	37.8	0.51	–
M890022	MC	1	319	68.97%	0.35	0.27	33.4	0.79	–
M890024	MC	1	281	91.07%	0.41	0.31	21.2	1.96	–
M890025	MC	1	319	67.71%	0.35	0.22	26.1	0.77	–
M890026	MC	1	319	43.57%	0.36	0.21	17.4	0.80	
M890029	MC	1	319	88.71%	0.35	0.25	32.3	1.13	–
M890030	MC	1	319	28.53%	0.37	0.22	32.7	0.89	

					$r_{score,\theta}$			IRT model includi	
Item	Format	Max. points	Number of obs.	Easiness	All items	TIMSS items only	Average response time [s]	Discrimination	diffi
M890039	MC	1	281	93.21%	0.44	0.23	43.6	2.11	−
LLM teachers' selection									
M090000	MC	1	304	83.88%	0.45	0.29	31.8	1.68	−
M090006	MC	1	304	93.09%	0.48	0.29	19.7	2.77	−
M090007	CR	1	304	21.05%	0.15	0.11	50.1	0.41	
M090012	MC	1	281	91.79%	0.24	0.13	33.6	0.94	−
M090013	CR	1	281	49.29%	0.44	0.24	46.7	1.04	
M090016	MC	1	281	60.36%	0.48	0.29	26.7	1.26	−
M090017	MC	1	281	9.29%	−0.11	−0.05	31.3	−0.29	−
M090042	CR	1	298	12.12%	0.47	0.26	64.4	1.88	
M090044	MC	1	299	18.12%	0.22	0.18	29.6	0.63	
M090047	CR	2	300	29.19%	0.46	0.29	184.8	0.99	
M090048	CR	1	298	0.34%	−0.03	−0.04	86.1	−0.47	−1
M090051	MC	1	299	65.10%	0.53	0.40	56.7	1.72	−
M090052	CR	1	300	79.19%	0.45	0.33	80.0	1.45	−
M090055	MC	1	304	91.45%	0.38	0.25	24.5	1.59	−
M090057	CR	1	304	31.91%	0.44	0.38	47.7	1.23	
M090059	MC	1	304	82.89%	0.46	0.28	28.2	1.60	−

8.2 Science items

					$r_{score,\theta}$			IRT model includin	
Item	Format	Max. points	Number of obs.	Easiness	All items	TIMSS items only	Average response time [s]	Discrimination	I diffic
TIMSS									
S041006	CR	2	305	30.94%	0.37	0.34	96.5	0.87	
S041010	MC	1	306	85.90%	0.35	0.32	19.3	1.01	−2
S041027	CR	1	294	94.90%	0.30	0.27	65.7	1.71	−2
S041032	CR	1	294	81.29%	0.43	0.44	61.9	1.36	−1
S041033	CR	1	299	16.11%	0.38	0.31	61.9	1.13	
S041043	CR	1	299	65.10%	0.38	0.42	49.3	0.94	−0
S041052	MC	1	305	66.12%	0.36	0.39	25.9	0.94	−0
S041068	CR	1	305	34.54%	0.37	0.46	49.8	0.95	0
S041070	MC	1	306	52.79%	0.30	0.37	31.2	0.74	−0
S041081	CR	2	305	39.93%	0.28	0.26	70.5	0.53	0
S041102	MC	1	305	78.62%	0.46	0.46	36.5	1.39	−1
S041107	CR	2	294	86.73%	0.53	0.61	42.5	2.12	−1
S041183	CR	2	306	46.39%	0.41	0.42	61.8	1.15	0
S041209	MC	1	299	64.77%	0.38	0.30	20.7	0.83	−0
S051037	CR	1	294	53.74%	0.45	0.44	63.2	1.08	−0
S051087	MC	1	306	81.31%	0.44	0.40	26.9	1.26	−1
S051105	MC	1	306	73.77%	0.40	0.44	29.1	1.13	−1
S051110	MC	1	294	61.22%	0.43	0.47	33.6	1.18	−0
S051111	MC	1	294	64.63%	0.14	0.20	24.4	0.38	−1
S051114	MC	1	299	56.71%	0.33	0.37	44.5	0.70	−0
S051121	MC	1	294	56.46%	0.34	0.29	45.7	0.68	−0
S051147	CR	1	294	27.21%	0.44	0.47	45.2	1.37	

					$r_{score,\theta}$			IRT model includin	
Item	Format	Max. points	Number of obs.	Easiness	All items	TIMSS items only	Average response time [s]	Discrimination	I diffic
S051164	CR	1	306	7.87%	0.39	0.27	67.9	2.00	
S051188	MC	1	294	51.02%	0.32	0.26	37.9	0.77	(
S051201	CR	1	306	37.38%	0.29	0.28	57.4	0.65	(
S061010	CR	1	306	58.69%	0.49	0.47	53.2	1.29	−(
S061028	MC	1	305	80.92%	0.33	0.38	18.8	1.05	−
S061031	CR	1	294	41.84%	0.45	0.44	44.9	1.22	(
S061034	CR	1	306	27.87%	0.40	0.46	50.1	1.23	(
S061060	CR	1	299	34.56%	0.33	0.29	68.0	0.80	(
S061065	MC	1	305	63.82%	0.31	0.35	23.1	0.83	−(
S061071	MC	1	299	77.18%	0.11	0.14	21.4	0.23	−5
S061098	MC	1	306	38.03%	0.23	0.12	33.6	0.38	
S061105	MC	1	305	65.46%	0.27	0.24	33.5	0.61	−
S061130	CR	1	305	21.05%	0.41	0.32	57.6	1.25	
S061138	CR	1	306	57.10%	0.35	0.43	77.0	0.92	−(
SP61011	CR	1	306	64.92%	0.56	0.63	67.5	1.81	−(
SP61044	CR	1	305	56.25%	0.34	0.40	48.2	0.72	−(
SP71017	MC	1	299	63.42%	0.36	0.41	53.8	0.89	−(
SP71054	CR	1	305	26.64%	0.37	0.42	66.1	1.03	
SP71072	MC	1	294	52.72%	0.36	0.42	44.9	0.85	−(
SP71077	CR	1	299	38.93%	0.57	0.61	61.7	1.79	(
SP71115	MC	1	299	56.71%	0.26	0.25	43.5	0.57	−(
SP71128	MC	1	294	78.23%	0.28	0.32	26.5	0.86	−
SP71140	MC	1	299	73.83%	0.40	0.44	24.5	1.31	−
SP71147	MC	1	305	69.08%	0.52	0.54	41.1	1.84	−(

					$r_{score,\theta}$			IRT model includin	
Item	Format	Max. points	Number of obs.	Easiness	All items	TIMSS items only	Average response time [s]	Discrimination	I diffic
SP71268	MC	1	299	46.31%	0.18	0.20	49.3	0.54	(
SP71402	MC	1	299	78.52%	0.37	0.36	22.2	0.95	−1
LLM simple setup									
S100002	MC	1	299	71.81%	0.42	0.31	33.0	0.91	−1
S100004	CR	1	299	2.69%	0.19	0.07	89.6	1.12	3
S100006	CR	0	299	0.00%			76.7		
S100008	MC	1	299	78.19%	0.44	0.25	26.9	1.44	−1
S100009	MC	1	299	22.15%	0.09	0.04	22.6	0.28	4
S100010	MC	1	299	96.64%	0.29	0.20	14.5	2.38	−2
S100012	CR	1	299	20.13%	0.36	0.24	71.6	1.08	1
S100014	MC	1	319	15.67%	0.02	0.04	27.8	0.16	10
S100015	MC	1	305	28.62%	0.15	0.16	58.2	0.36	2
S100016	MC	1	305	72.04%	0.28	0.18	30.7	0.71	−1
S100020	MC	1	319	56.11%	0.41	0.27	38.6	0.98	−0
S100023	CR	1	319	8.46%	0.34	0.21	84.6	1.55	2
S100024	CR	1	305	15.79%	0.21	0.11	79.9	0.61	2
S100029	MC	1	305	73.36%	0.38	0.28	28.7	1.01	−1
S100032	CR	1	319	4.73%	0.28	0.15	52.2	1.59	2
S100039	CR	2	304	9.57%	0.49	0.21	120.0	1.82	1
LLM complex setup									
S890002	CR	1	281	9.29%	0.30	0.13	78.9	0.95	2
S890004	CR	1	281	70.71%	0.30	0.23	40.8	0.71	−1
S890005	CR	1	319	4.08%	0.27	0.23	126.4	1.67	2
S890006	CR	2	281	5.54%	0.31	0.24	127.5	1.32	3

					$r_{score,\theta}$			IRT model includin	
Item	Format	Max. points	Number of obs.	Easiness	All items	TIMSS items only	Average response time [s]	Discrimination	I diffic
S890007	CR	1	281	12.86%	0.33	0.18	79.7	1.13	2
S890009	CR	1	281	55.00%	0.48	0.32	56.2	1.07	−(
S890010	MC	1	319	69.28%	0.36	0.24	36.6	0.86	−1
S890011	MC	1	319	47.65%	0.28	0.18	36.7	0.65	(
S890016	MC	1	319	48.28%	0.16	0.06	28.1	0.32	(
S890017	CR	2	319	2.51%	0.27	0.12	85.7	1.45	3
S890021	MC	1	319	76.18%	0.06	0.00	34.7	0.23	−5
S890022	MC	1	281	86.43%	0.34	0.21	17.5	1.12	−2
S890023	MC	1	319	64.89%	0.35	0.25	29.0	0.86	−(
S890024	MC	1	319	46.71%	0.20	0.17	23.1	0.45	(
S890026	MC	1	281	48.93%	0.30	0.12	26.3	0.65	(
S890038	MC	1	281	82.86%	0.35	0.19	17.5	1.05	−1
LLM teachers' selection									
S090001	CR	1	299	83.56%	0.43	0.25	65.4	1.45	−1
S090015	MC	1	304	45.39%	0.30	0.16	22.3	0.75	(
S090018	MC	1	304	71.05%	0.37	0.26	18.2	0.95	−1
S090020	MC	1	304	64.80%	0.48	0.34	27.7	1.26	−(
S090040	MC	1	299	87.92%	0.34	0.17	25.7	1.18	−2
S090041	MC	1	299	79.19%	0.32	0.21	33.9	1.02	−1
S090048	MC	1	299	36.24%	0.05	0.03	31.0	0.21	2
S090051	MC	1	304	59.87%	0.20	0.09	19.2	0.44	−(
S090052	MC	1	299	94.63%	0.45	0.28	13.4	3.10	−1
S090053	MC	1	304	65.46%	0.38	0.25	24.0	0.90	−(
S090058	MC	1	281	21.79%	−0.08	−0.10	22.8	−0.19	−(

					$r_{score,\theta}$			IRT model includin	
Item	Format	Max. points	Number of obs.	Easiness	All items	TIMSS items only	Average response time [s]	Discrimination	I diffic
S090064	MC	1	281	66.07%	0.48	0.34	29.9	1.51	−(
S090075	MC	1	281	67.50%	0.30	0.24	15.7	0.79	−1
S090077	CR	1	304	42.43%	0.26	0.08	62.0	0.49	(
S090080	MC	1	281	75.00%	0.37	0.23	18.1	1.04	−1
S090084	MC	1	304	64.47%	0.31	0.17	22.3	0.79	−(

9 Problem summaries

In tables below numer of items with probelmatic properties were summarised. Criteria for each of the described problems were:

- *Negative discrimination*: a ≤ 0
- *Low discrimination*: a <0.5
- *Too easy*: b < -2
- *Too difficult*: b > 2
- *Gender-DIF*: gender-DIF detected (statistically significant at the 0.05 level in logistic regression DIF analysis)
- *Poor item fit*: item RMSE statistically significantly different from 0 at the 0.05 level
- *Local item dependence*: local dependency with some other item detected
- *Non-functional distractors*: only for MC items with easiness < 85%: at least one distractor chosen by less than 5% of respondents
- *Total*: number of items meeting at least one of the previous criteria

9.1 Math items

	N	Negative discrimination	Low discrimination	Too easy	Too difficult	Gender-DIF	Poor item fit	Local item dependence	fun dist
TIMSS									
Overall	47	0	3	1	1	0	2	0	
CR	14	0	0	0	1	0	1	0	
MC	33	0	3	1	0	0	1	0	
LLM									

	N	Negative discrimination	Low discrimination	Too easy	Too difficult	Gender-DIF	Poor item fit	Local item dependence	fun dist
Overall	48	3	2	9	5	0	7	0	
CR	23	1	2	2	4	0	3	0	
MC	25	2	0	7	1	0	4	0	
LLM simple setup									
Overall	16	1	0	2	1	0	2	0	
CR	8	0	0	0	1	0	0	0	
MC	8	1	0	2	0	0	2	0	
LLM complex setup									
Overall	16	0	1	3	2	0	1	0	
CR	8	0	1	1	2	0	0	0	
MC	8	0	0	2	0	0	1	0	
LLM teachers' selection									
Overall	16	2	1	4	2	0	4	0	
CR	7	1	1	1	1	0	3	0	
MC	9	1	0	3	1	0	1	0	

9.2 Science items

	N	Negative discrimination	Low discrimination	Too easy	Too difficult	Gender-DIF	Poor item fit	Local item dependence	fun dist
TIMSS									
Overall	48	0	3	3	0	2	3	0	
CR	23	0	0	1	0	1	2	0	
MC	25	0	3	2	0	1	1	0	
LLM									

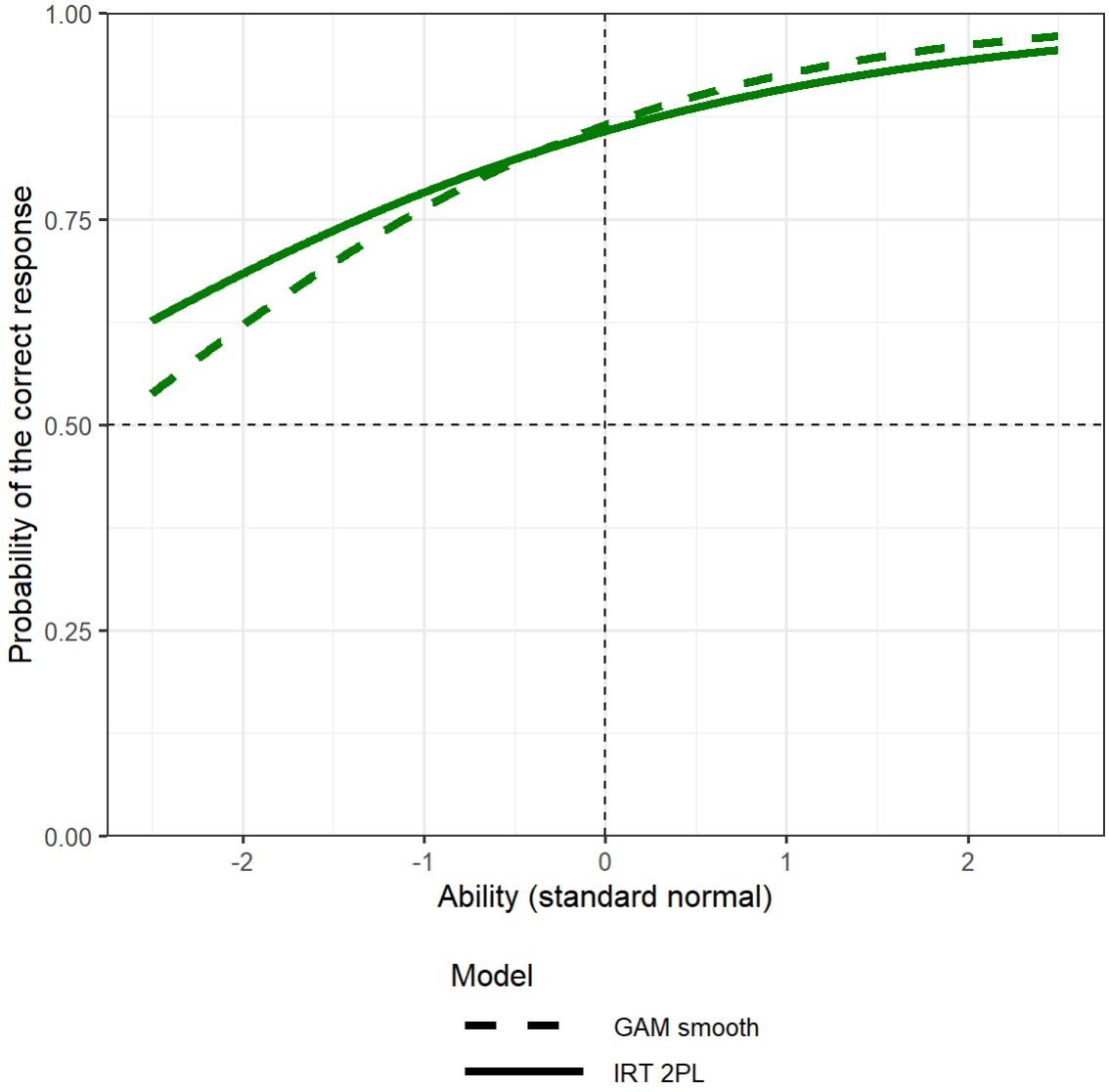
	N	Negative discrimination	Low discrimination	Too easy	Too difficult	Gender-DIF	Poor item fit	Local item dependence	fun dist
Overall	48	1	9	5	13	2	4	0	
CR	16	0	1	0	9	1	1	0	
MC	32	1	8	5	4	1	3	0	
LLM simple setup									
Overall	16	0	3	1	7	0	2	0	
CR	7	0	0	0	4	0	1	0	
MC	9	0	3	1	3	0	1	0	
LLM complex setup									
Overall	16	0	3	2	5	1	0	0	
CR	7	0	0	0	5	1	0	0	
MC	9	0	3	2	0	0	0	0	
LLM teachers' selection									
Overall	16	1	3	2	1	1	2	0	
CR	2	0	1	0	0	0	0	0	
MC	14	1	2	2	1	1	2	0	

10 Individual items

10.1 Math items

10.1.1 M890021 (LLM)

10.1.1.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	84.64%	319	LLM complex setup	MC	0.23	0.10	0.51	−3.51	0.00

10.1.1.2 Gender DIF:

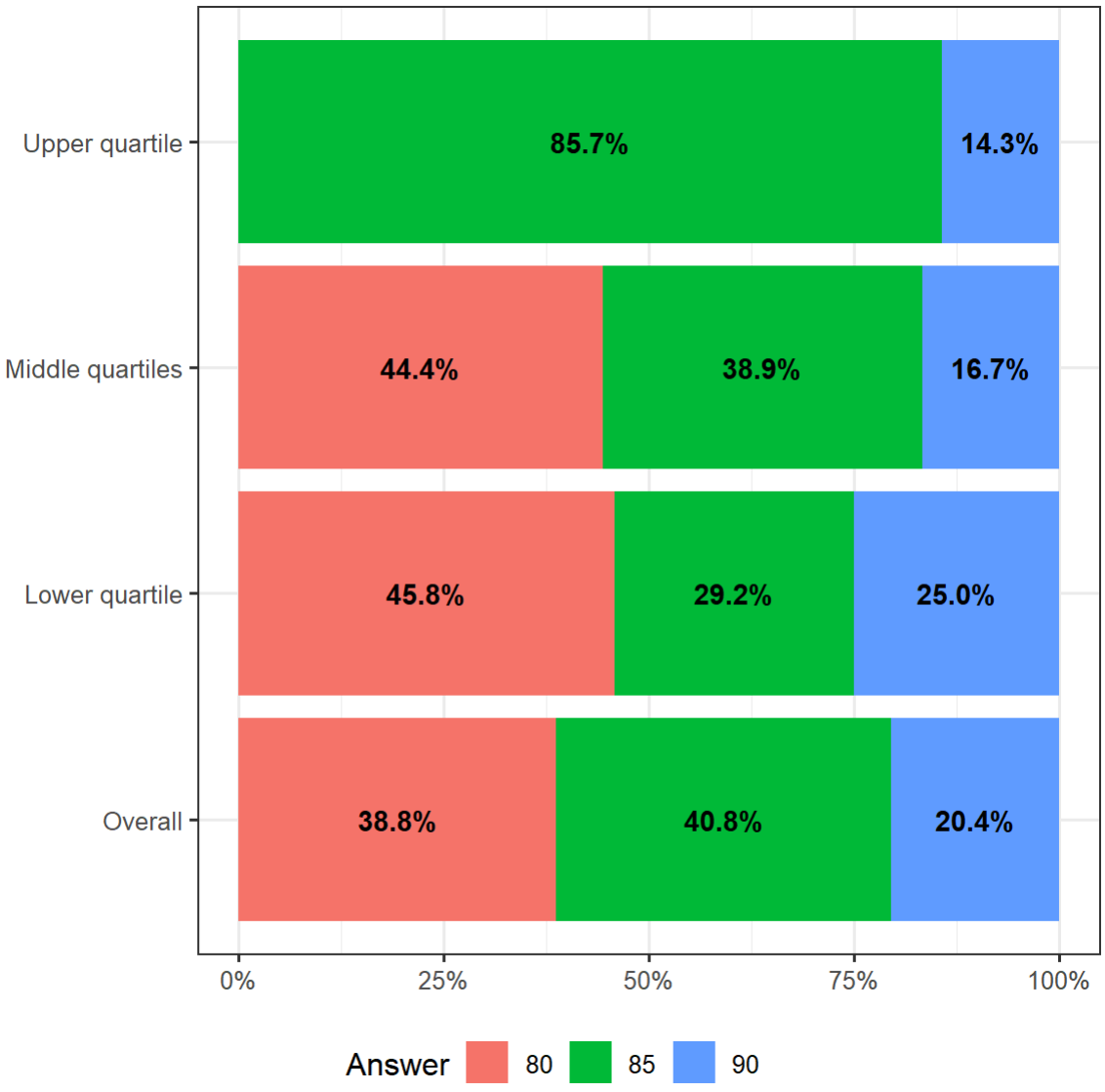
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.16	0.560	1.000	0.005	N/A

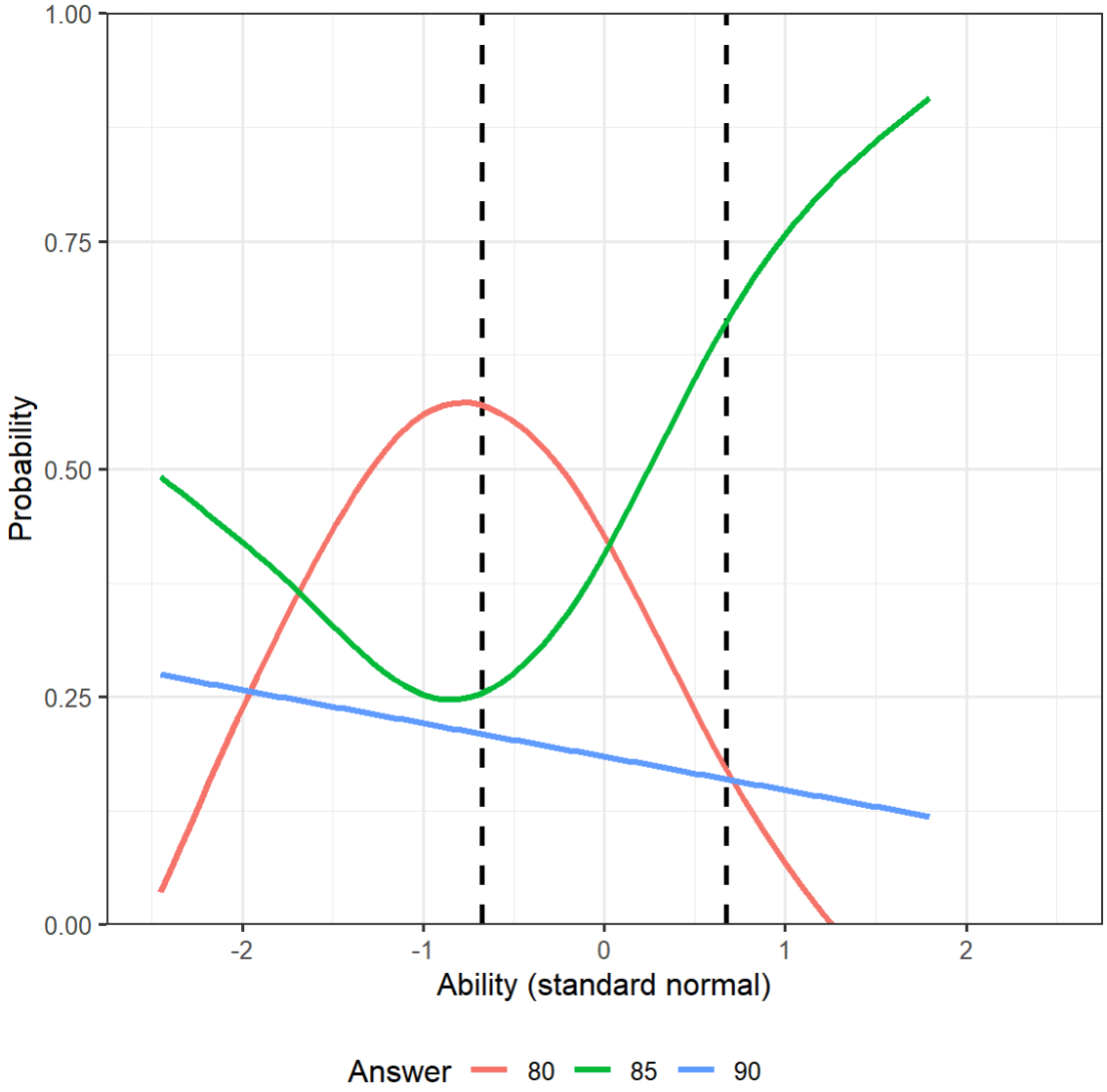
10.1.1.3 Frequency of answers:

Answer	Frequency	Percent
75 (correct)	270	84.6%
80	19	6.0%
85	20	6.3%
90	10	3.1%
Overall	319	100.0%

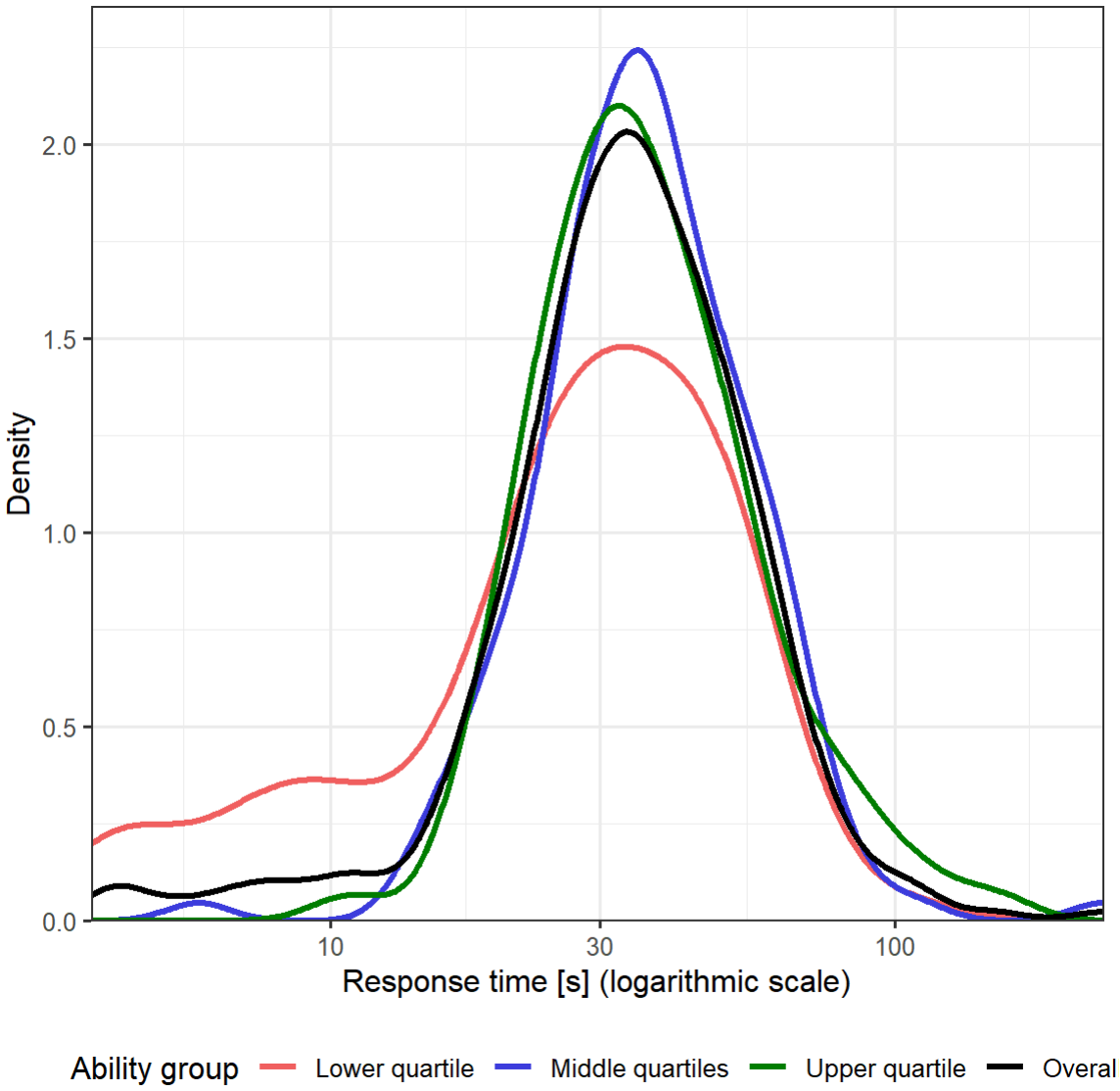
10.1.1.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
80	11	8	0	19	45.8%	44.4%	0.0%	38.8%
85	7	7	6	20	29.2%	38.9%	85.7%	40.8%
90	6	3	1	10	25.0%	16.7%	14.3%	20.4%
Overall	24	18	7	49	100.0%	100.0%	100.0%	100.0%


Upper quartile
Middle quartiles
Lower quartile
Overall
85.7%
14.3%
44.4%
38.9%
16.7%
45.8%
29.2%
25.0%
38.8%
40.8%
20.4%
0%
25%
50%
75%
100%
Answer
80
85
90



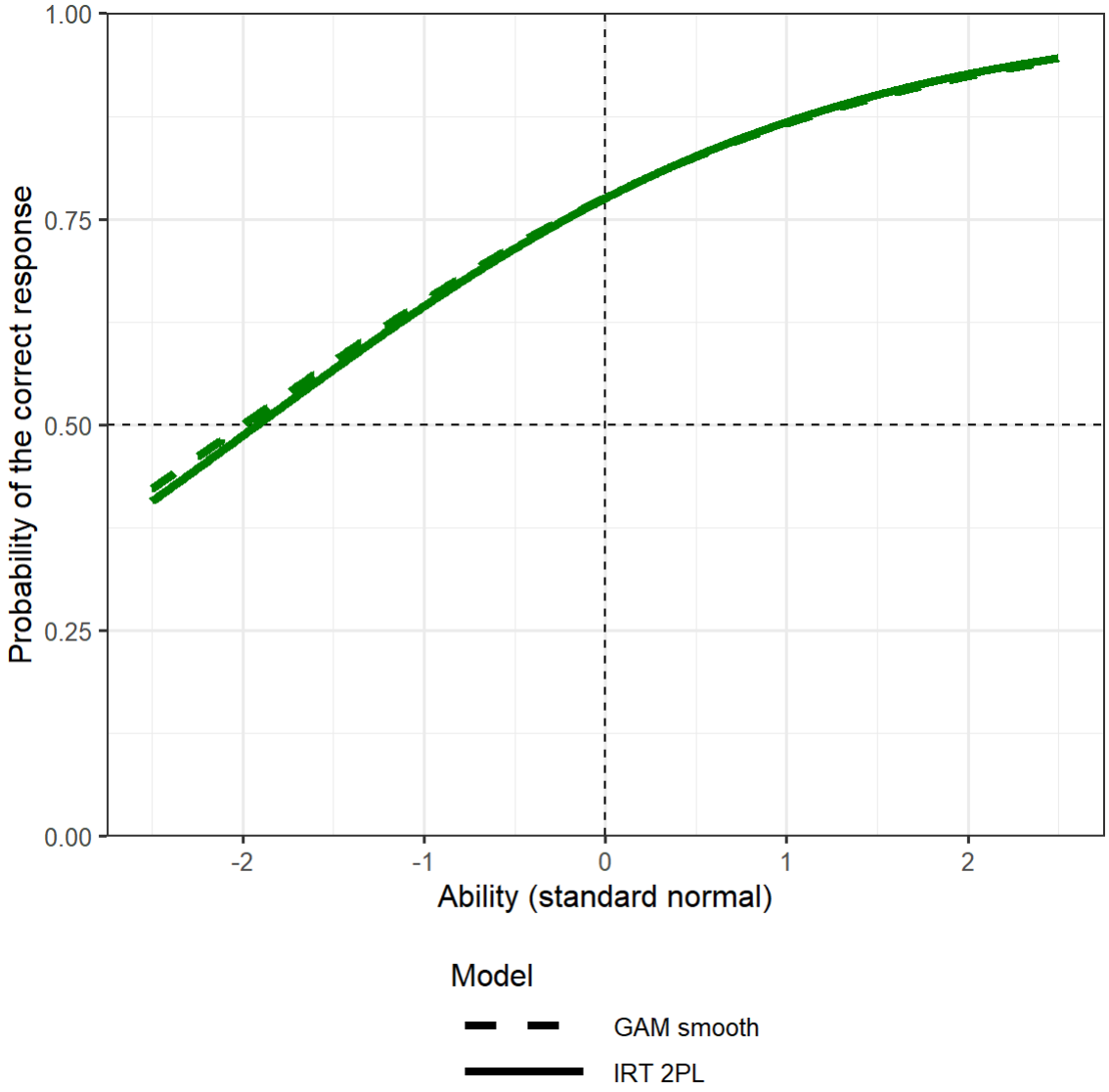
10.1.1.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.8	18.8	29.6	43.5	103.0	3.2	0.7
Middle quartiles	5.8	27.8	35.7	47.7	234.0	3.6	0.4
Upper quartile	10.8	26.2	34.6	47.0	150.3	3.6	0.5
Overall	3.8	25.5	34.4	46.1	234.0	3.5	0.6

10.1.2 M041135 (TIMSS)

10.1.2.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	76.17%	299	TIMSS	MC	0.25	0.27	0.65	−1.92	0.026

10.1.2.2 Gender DIF:

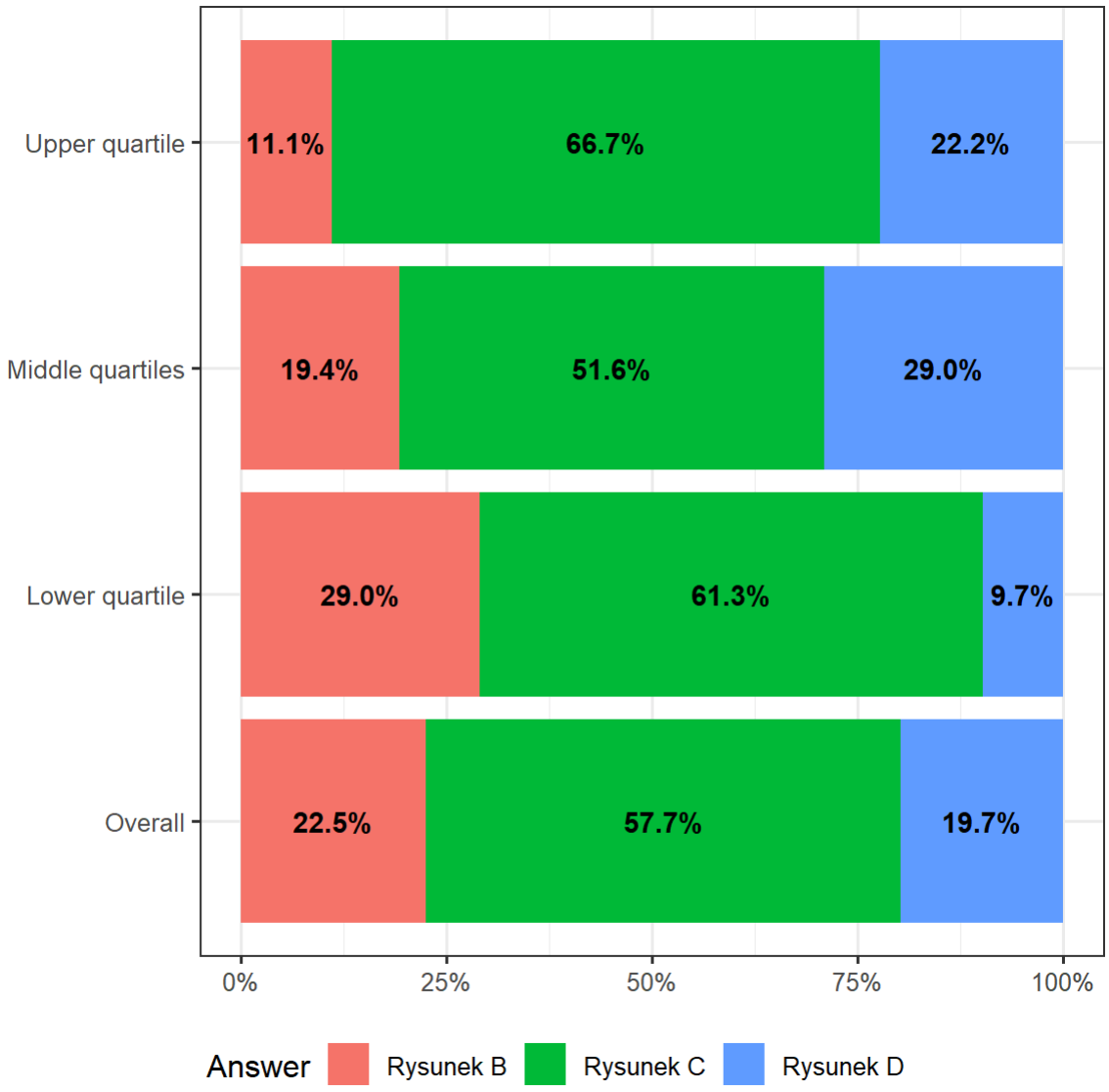
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
12.53	0.002	0.181	0.043	N/A

10.1.2.3 Frequency of answers:

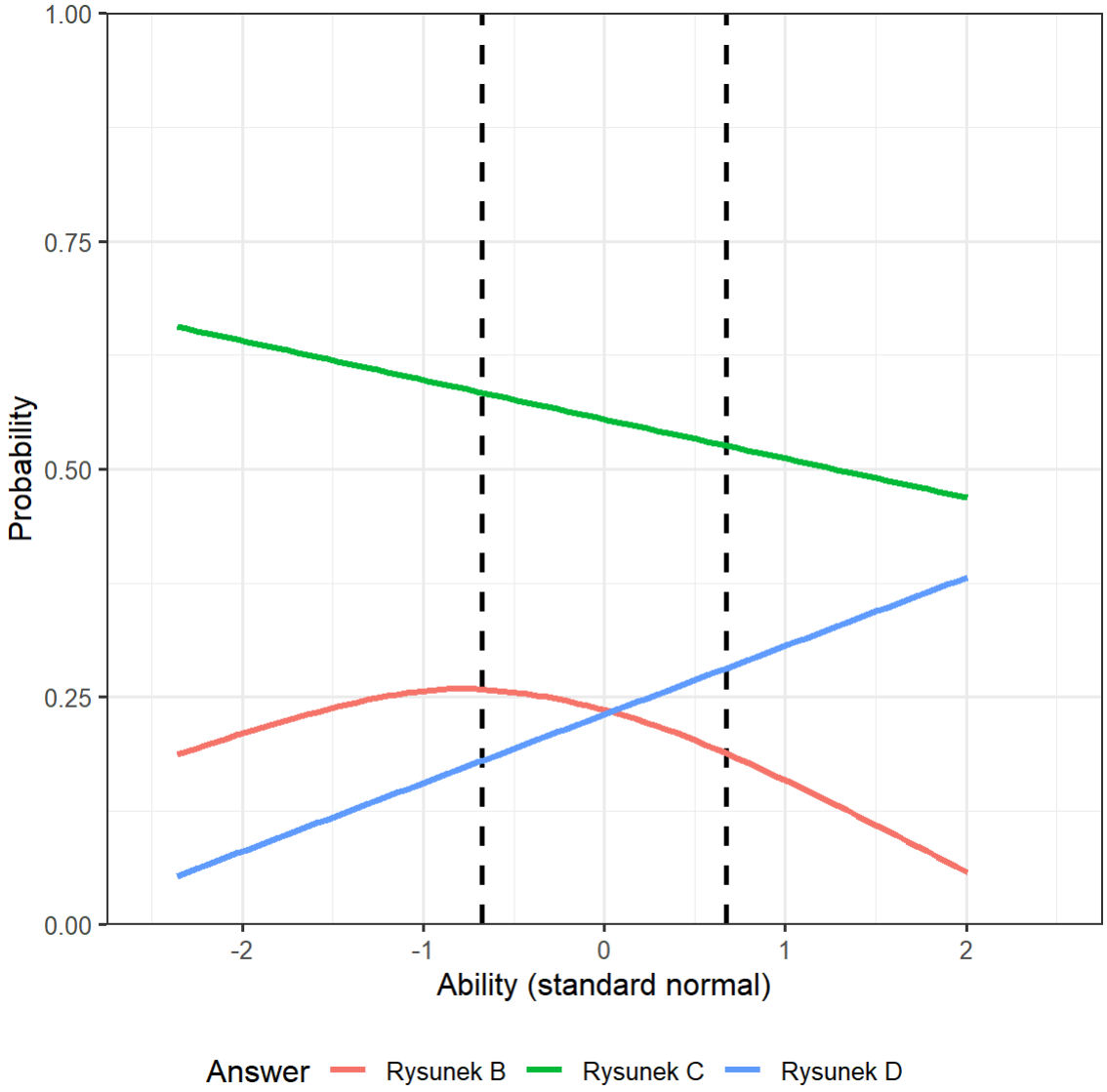
Answer	Frequency	Percent
Rysunek A (correct)	227	76.2%
Rysunek B	16	5.4%
Rysunek C	41	13.8%
Rysunek D	14	4.7%
Overall	298	100.0%

10.1.2.4 Distractors functioning:

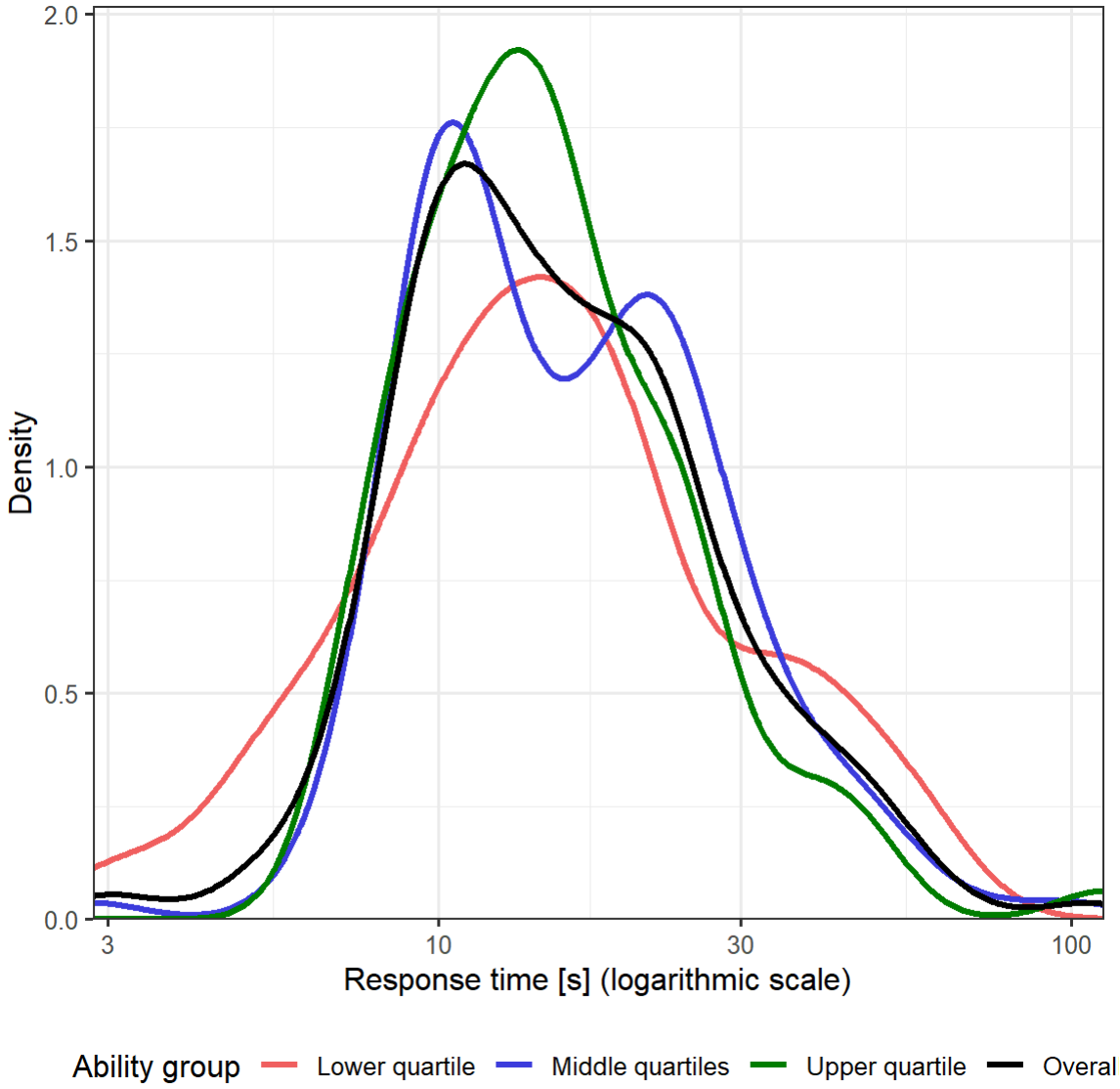
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Rysunek B	9	6	1	16	29.0%	19.4%	11.1%	22.5%
Rysunek C	19	16	6	41	61.3%	51.6%	66.7%	57.7%
Rysunek D	3	9	2	14	9.7%	29.0%	22.2%	19.7%
Overall	31	31	9	71	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



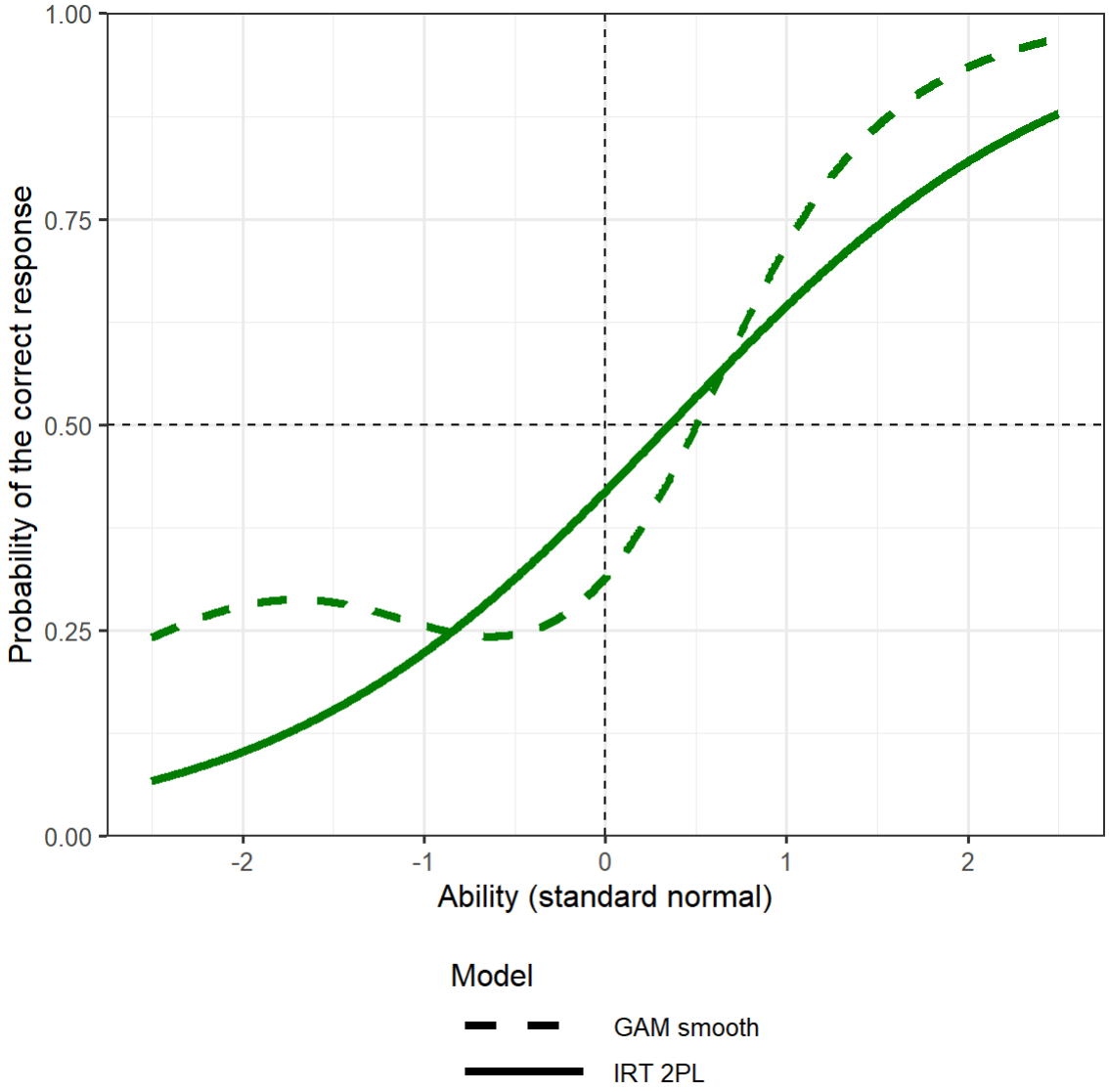
10.1.2.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.9	9.8	14.3	21.2	58.4	2.7	0.7
Middle quartiles	2.8	10.1	15.1	22.7	101.2	2.8	0.6
Upper quartile	7.1	10.8	14.1	20.1	112.3	2.7	0.5
Overall	2.8	10.2	14.8	22.1	112.3	2.7	0.6

10.1.3 M041268 (TIMSS)

10.1.3.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	43.96%	299	TIMSS	MC	0.40	0.38	0.92	0.36	0.035

10.1.3.2 Gender DIF:

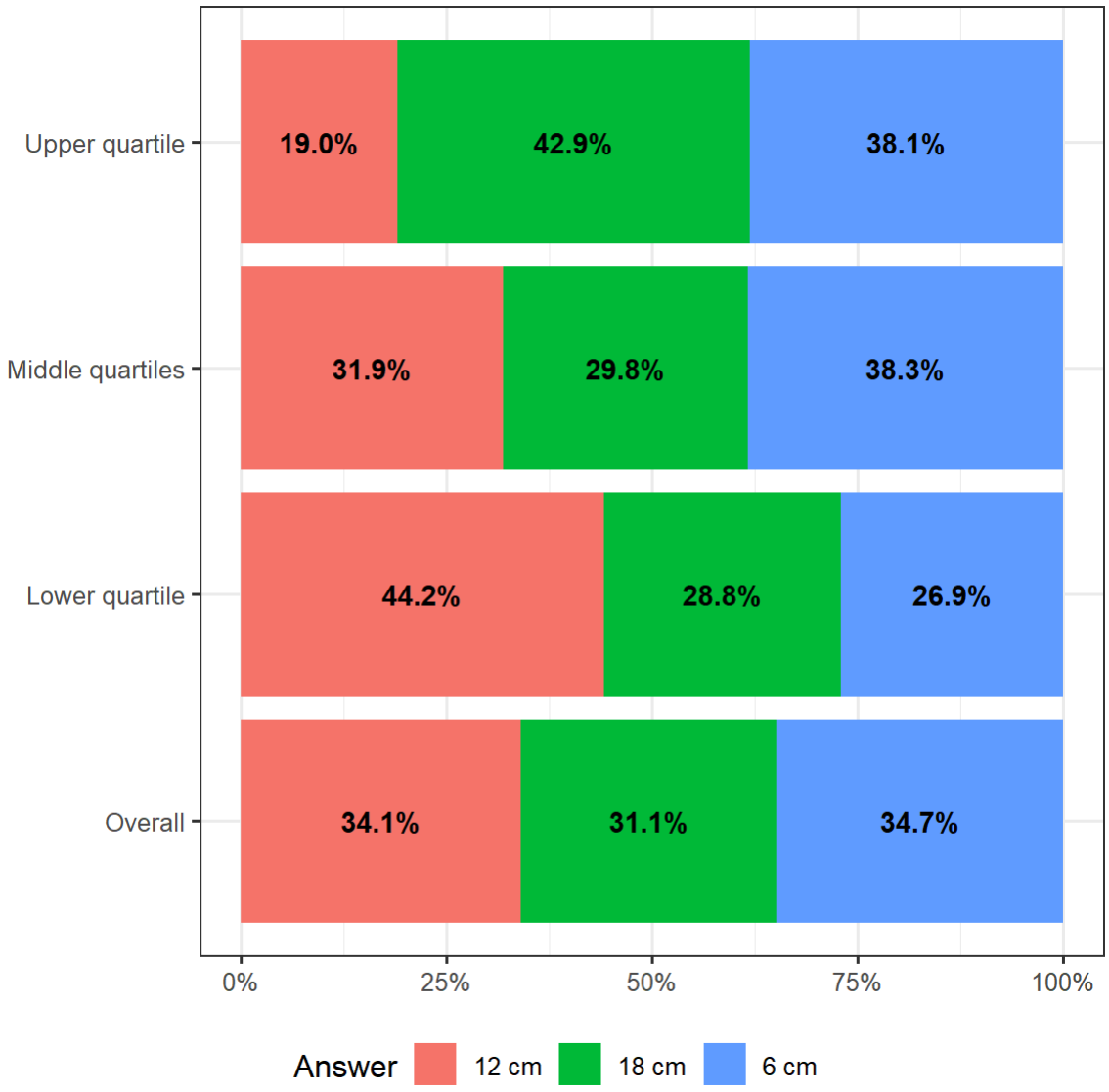
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.27	0.321	1.000	0.006	N/A

10.1.3.3 Frequency of answers:

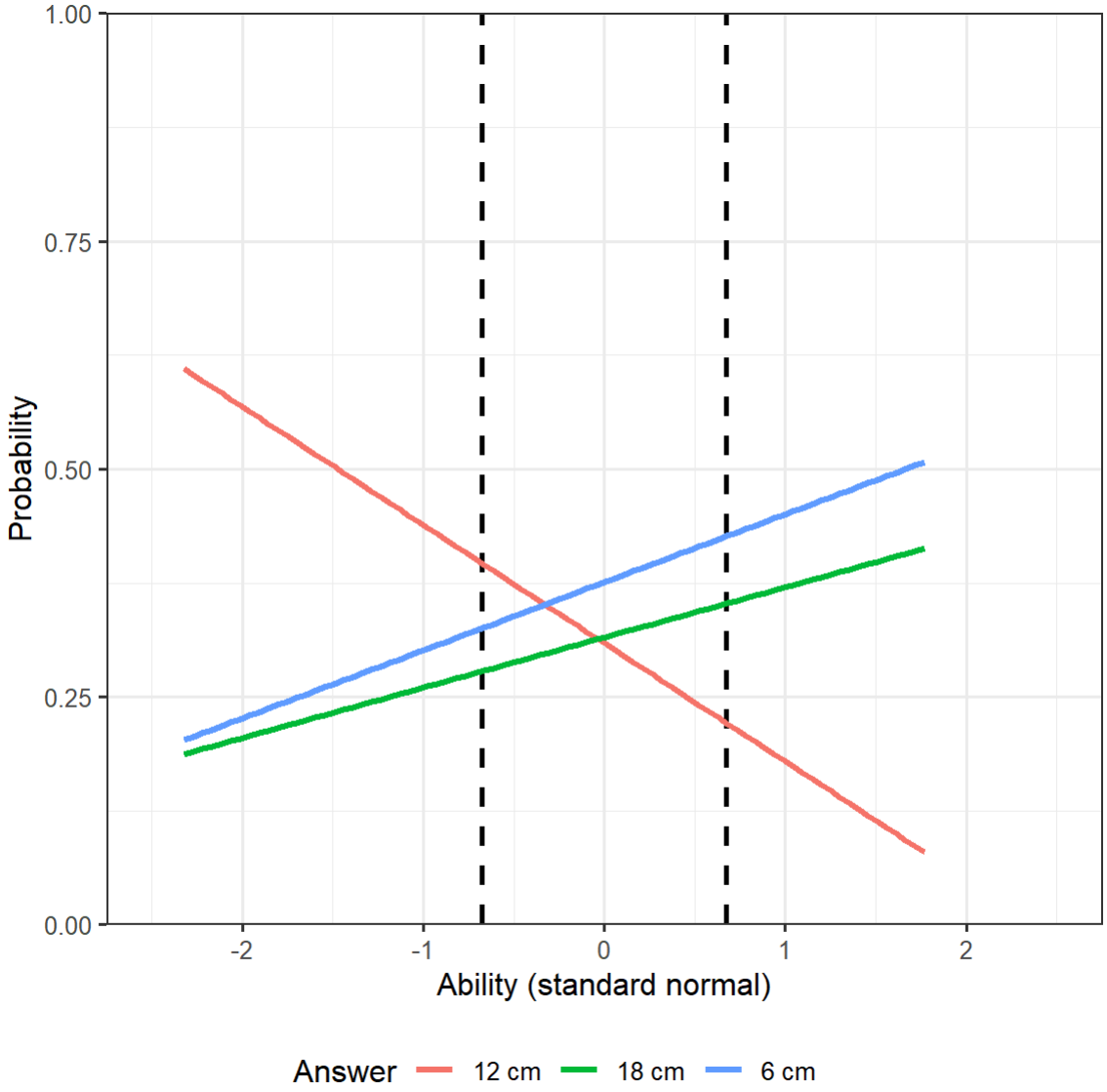
Answer	Frequency	Percent
12 cm	57	19.1%
18 cm	52	17.4%
6 cm	58	19.5%
9 cm (correct)	131	44.0%
Overall	298	100.0%

10.1.3.4 Distractors functioning:

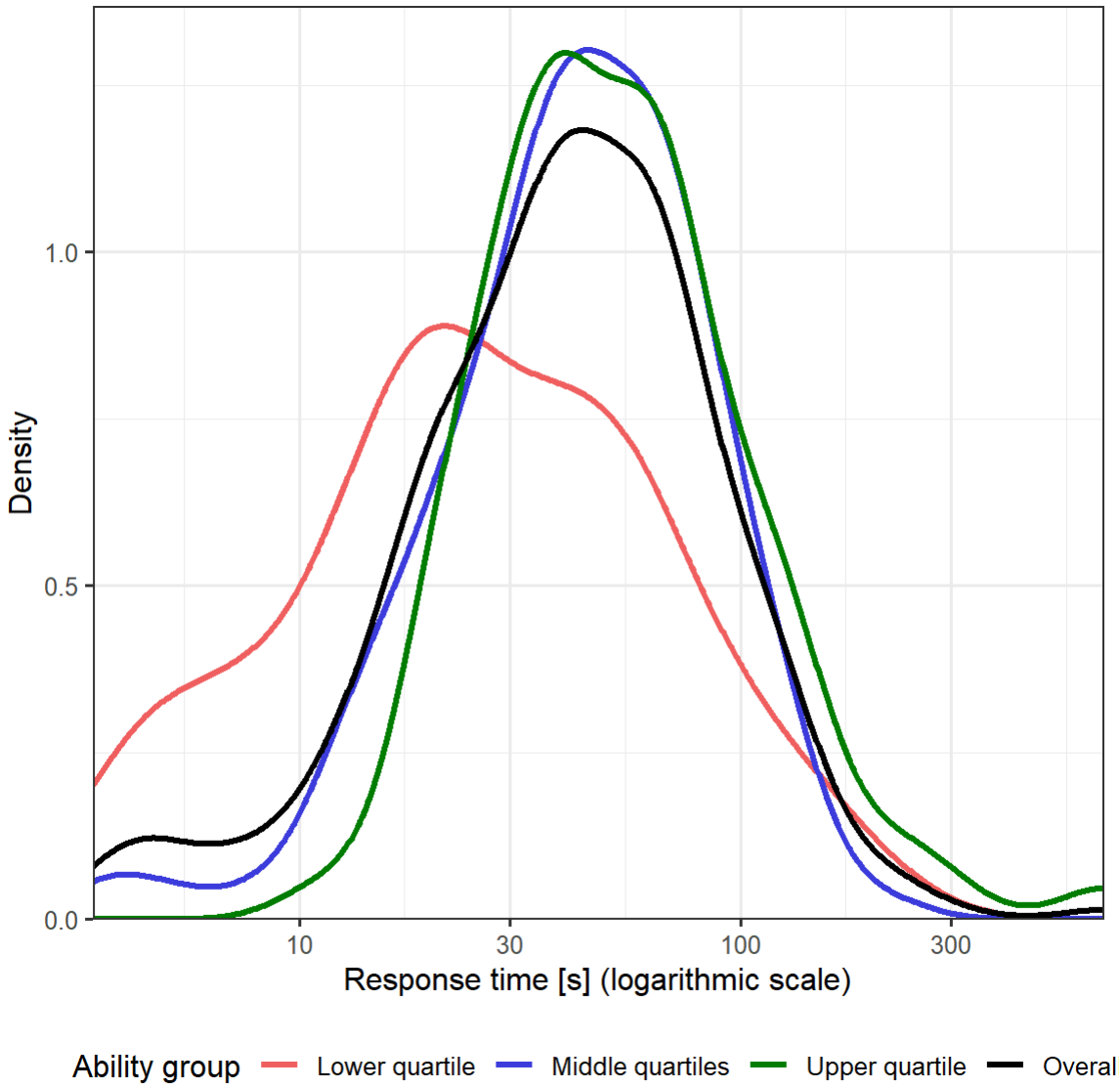
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
12 cm	23	30	4	57	44.2%	31.9%	19.0%	34.1%
18 cm	15	28	9	52	28.8%	29.8%	42.9%	31.1%
6 cm	14	36	8	58	26.9%	38.3%	38.1%	34.7%
Overall	52	94	21	167	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



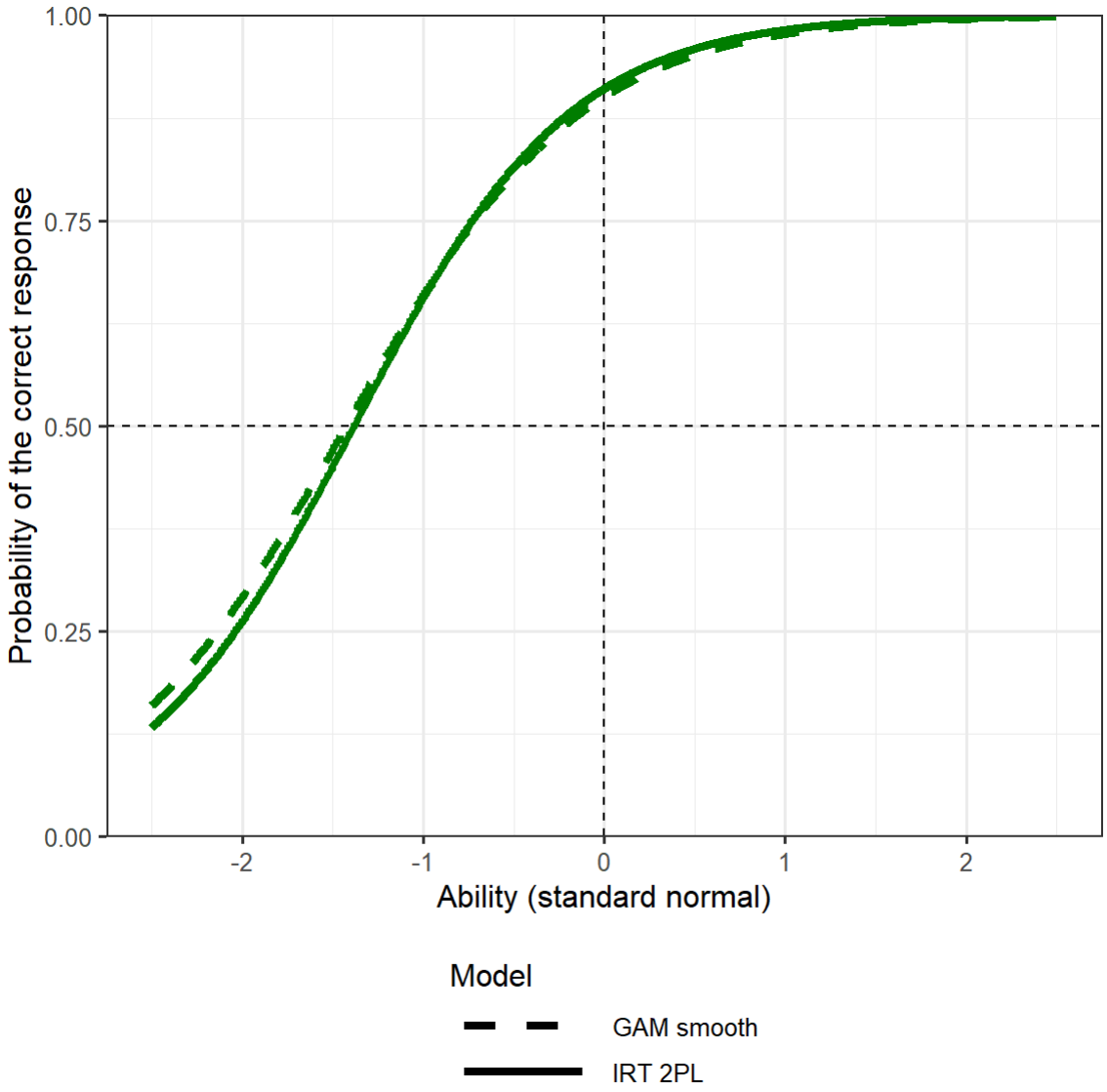
10.1.3.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	14.6	24.1	51.0	196.9	3.2	1.0
Middle quartiles	3.4	27.1	42.8	67.4	211.2	3.7	0.7
Upper quartile	11.2	32.6	49.8	75.7	663.8	4.0	0.7
Overall	3.4	23.8	41.4	66.7	663.8	3.7	0.8

10.1.4 M090000 (LLM)

10.1.4.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	83.88%	304	LLM teachers' selection	MC	0.45	0.29	1.68	−1.38	0.0(

10.1.4.2 Gender DIF:

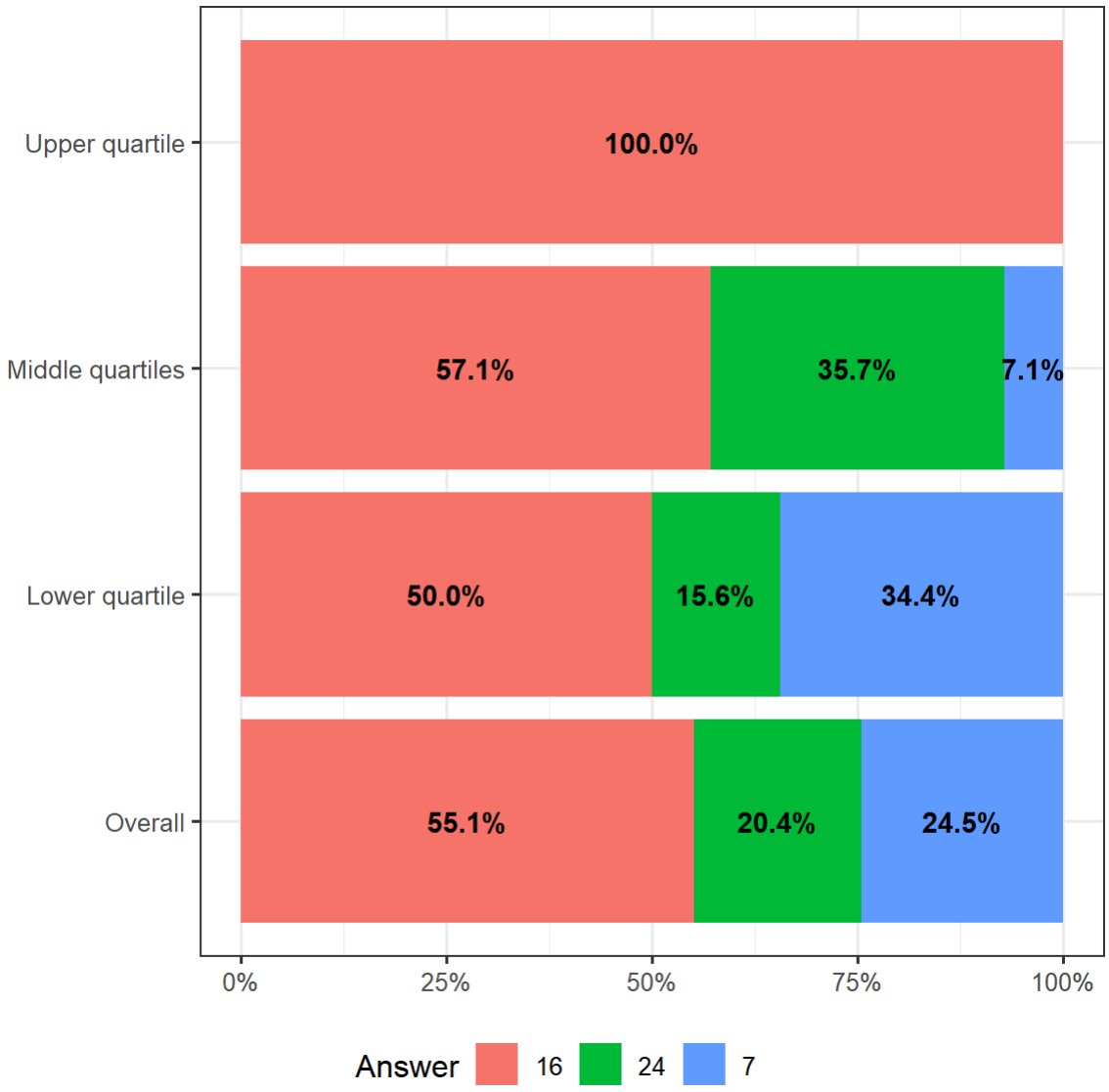
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
7.30	0.026	1.000	0.028	N/A

10.1.4.3 Frequency of answers:

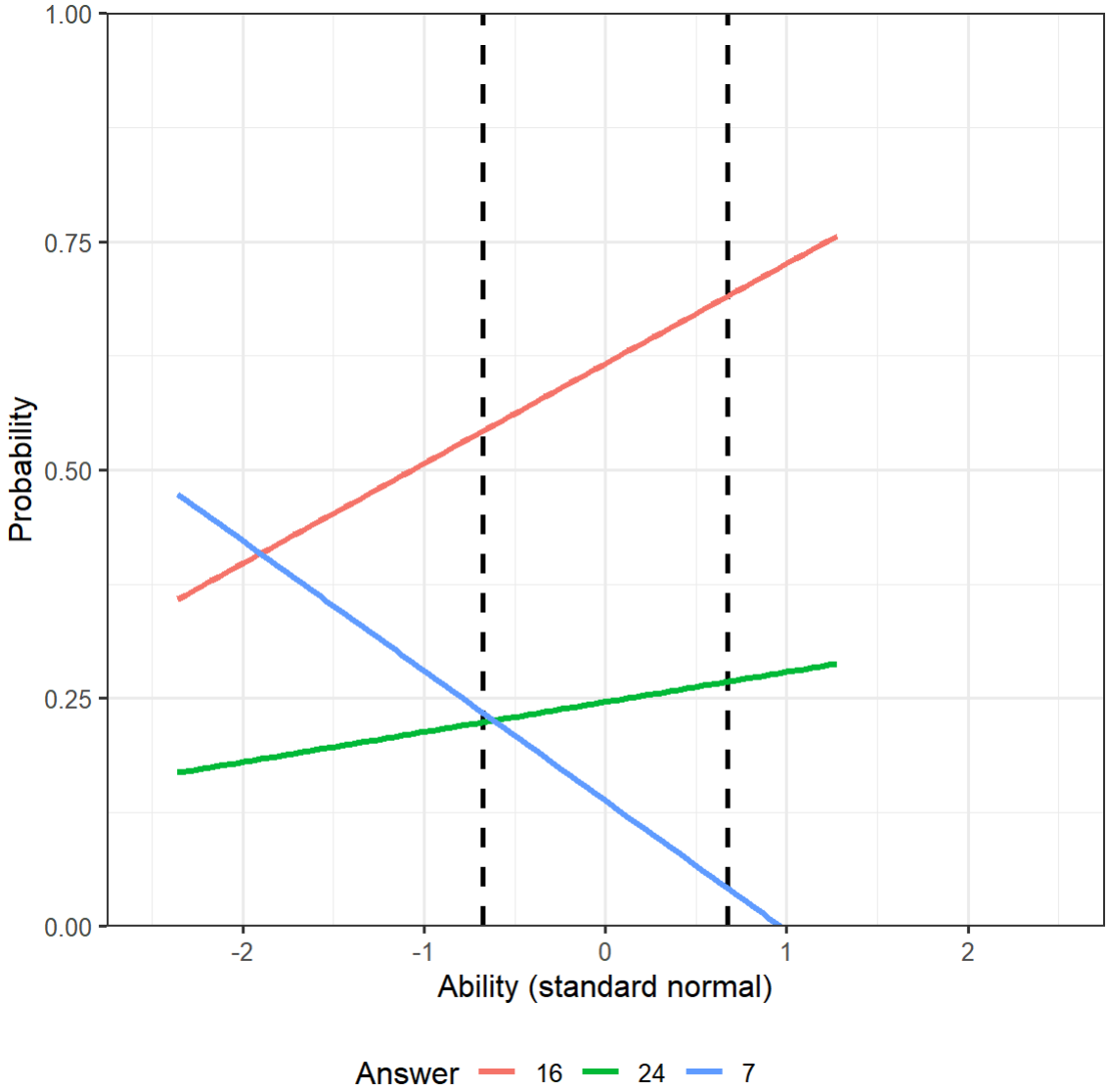
Answer	Frequency	Percent
7	12	3.9%
12 (correct)	255	83.9%
16	27	8.9%
24	10	3.3%
Overall	304	100.0%

10.1.4.4 Distractors functioning:

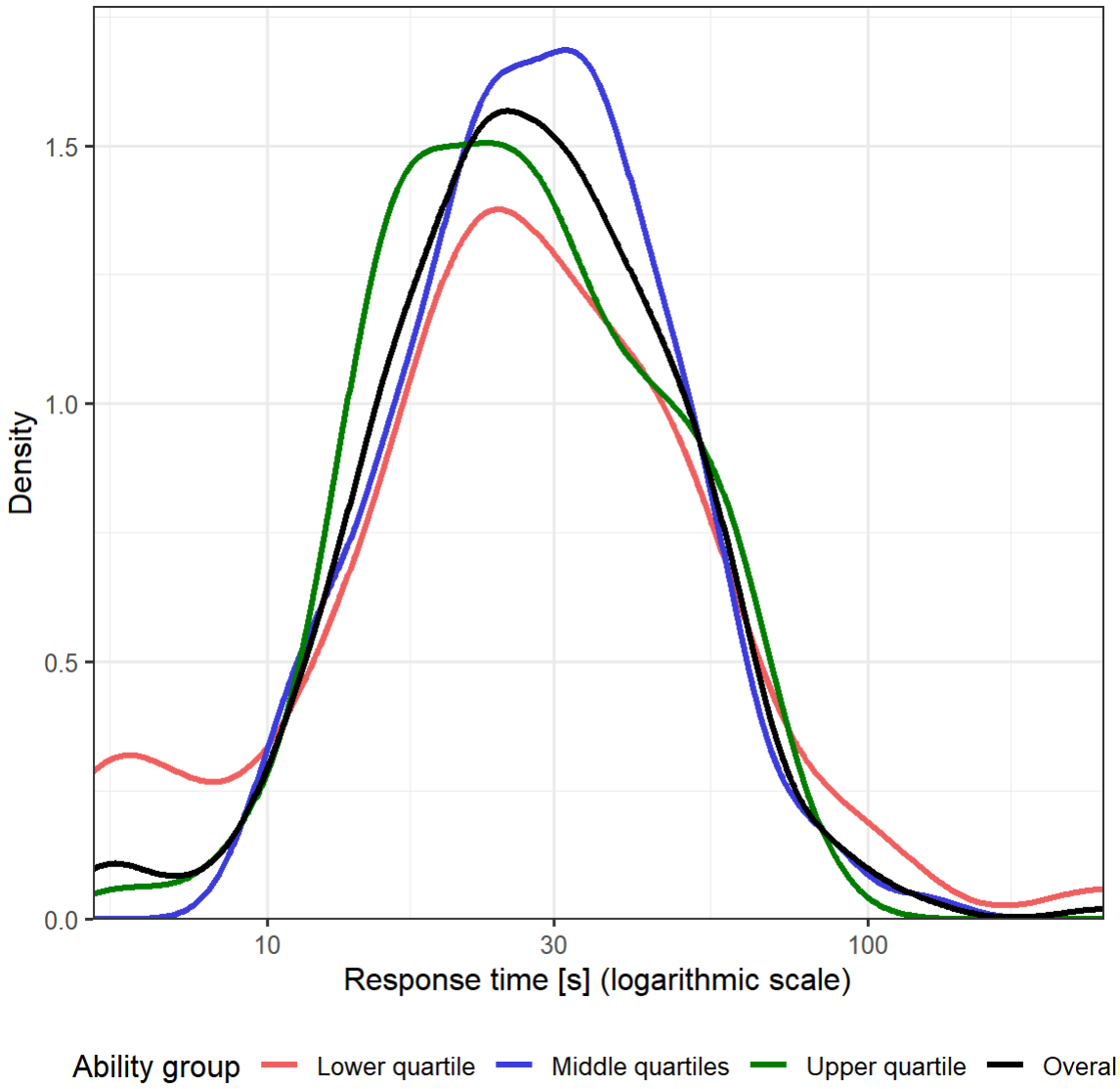
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
16	16	8	3	27	50.0%	57.1%	100.0%	55.1%
24	5	5	0	10	15.6%	35.7%	0.0%	20.4%
7	11	1	0	12	34.4%	7.1%	0.0%	24.5%
Overall	32	14	3	49	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



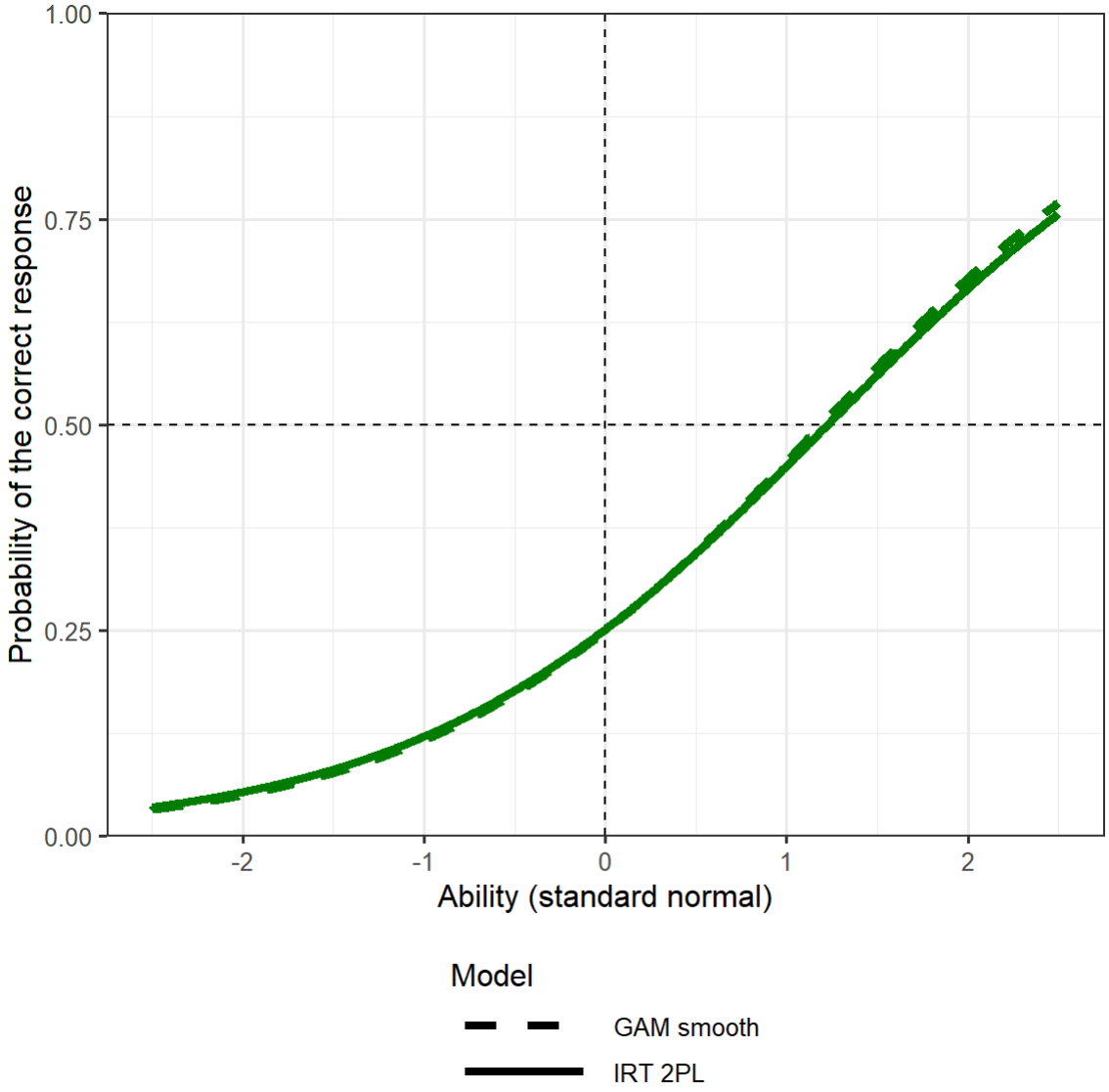
10.1.4.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.1	18.5	25.3	41.3	246.9	3.2	0.7
Middle quartiles	9.8	20.7	28.1	38.8	124.0	3.3	0.5
Upper quartile	5.7	17.4	25.7	39.8	78.4	3.3	0.5
Overall	5.1	18.9	27.4	39.4	246.9	3.3	0.6

10.1.5 M890030 (LLM)

10.1.5.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	28.53%	319	LLM complex setup	MC	0.37	0.22	0.89	1.23	0.01

10.1.5.2 Gender DIF:

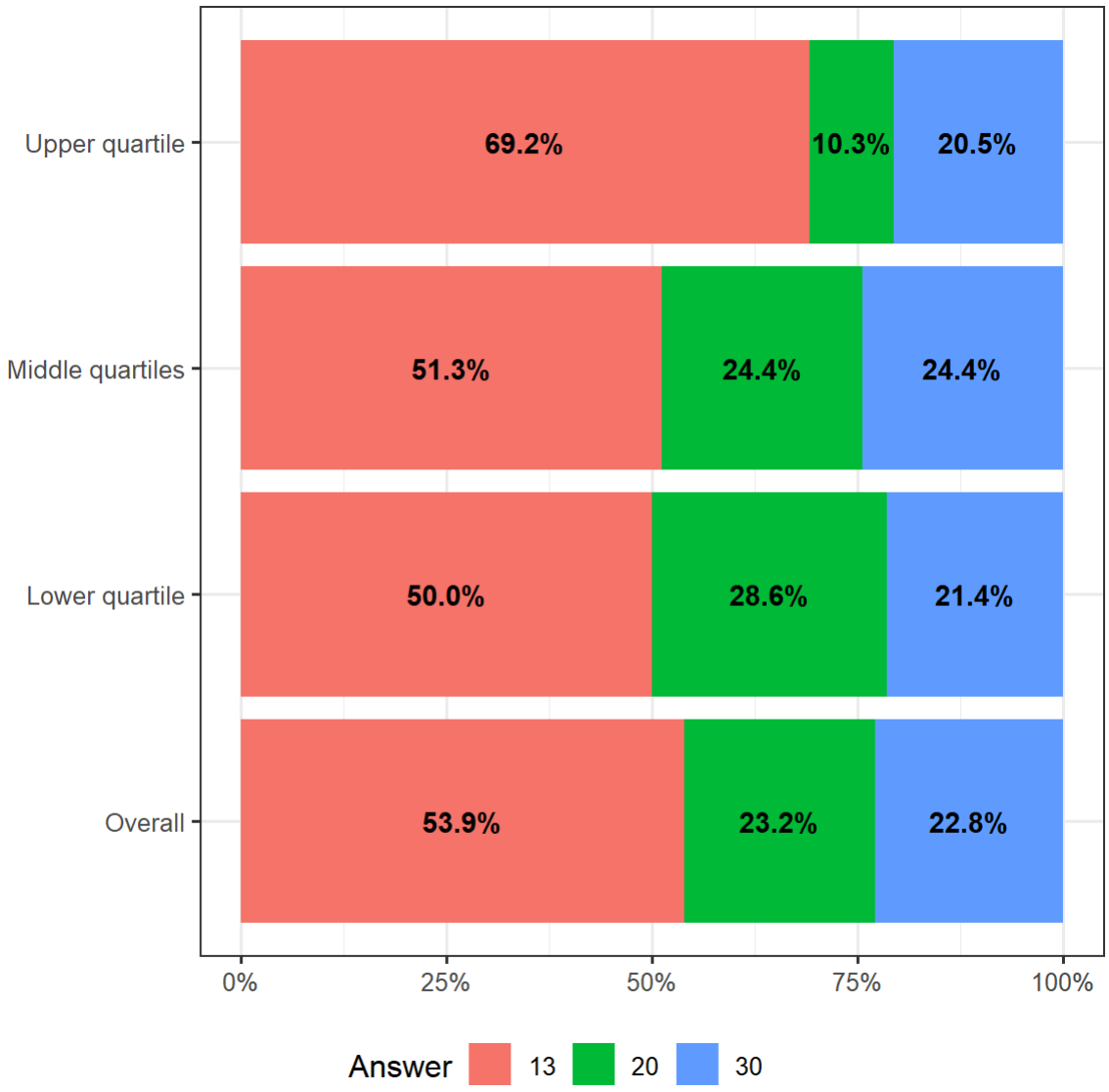
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.56	0.459	1.000	0.005	N/A

10.1.5.3 Frequency of answers:

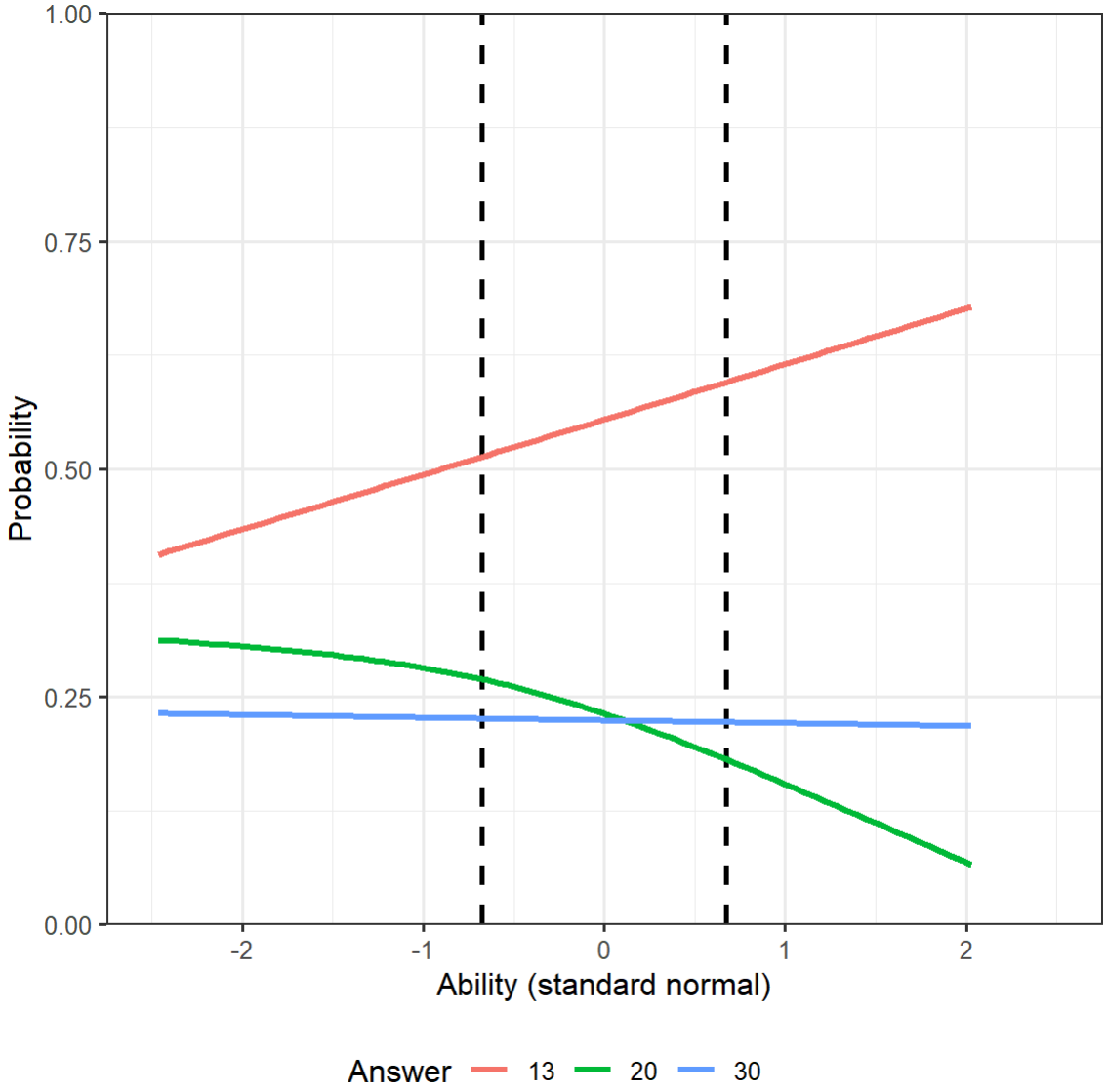
Answer	Frequency	Percent
13	123	38.6%
20	53	16.6%
30	52	16.3%
40 (correct)	91	28.5%
Overall	319	100.0%

10.1.5.4 Distractors functioning:

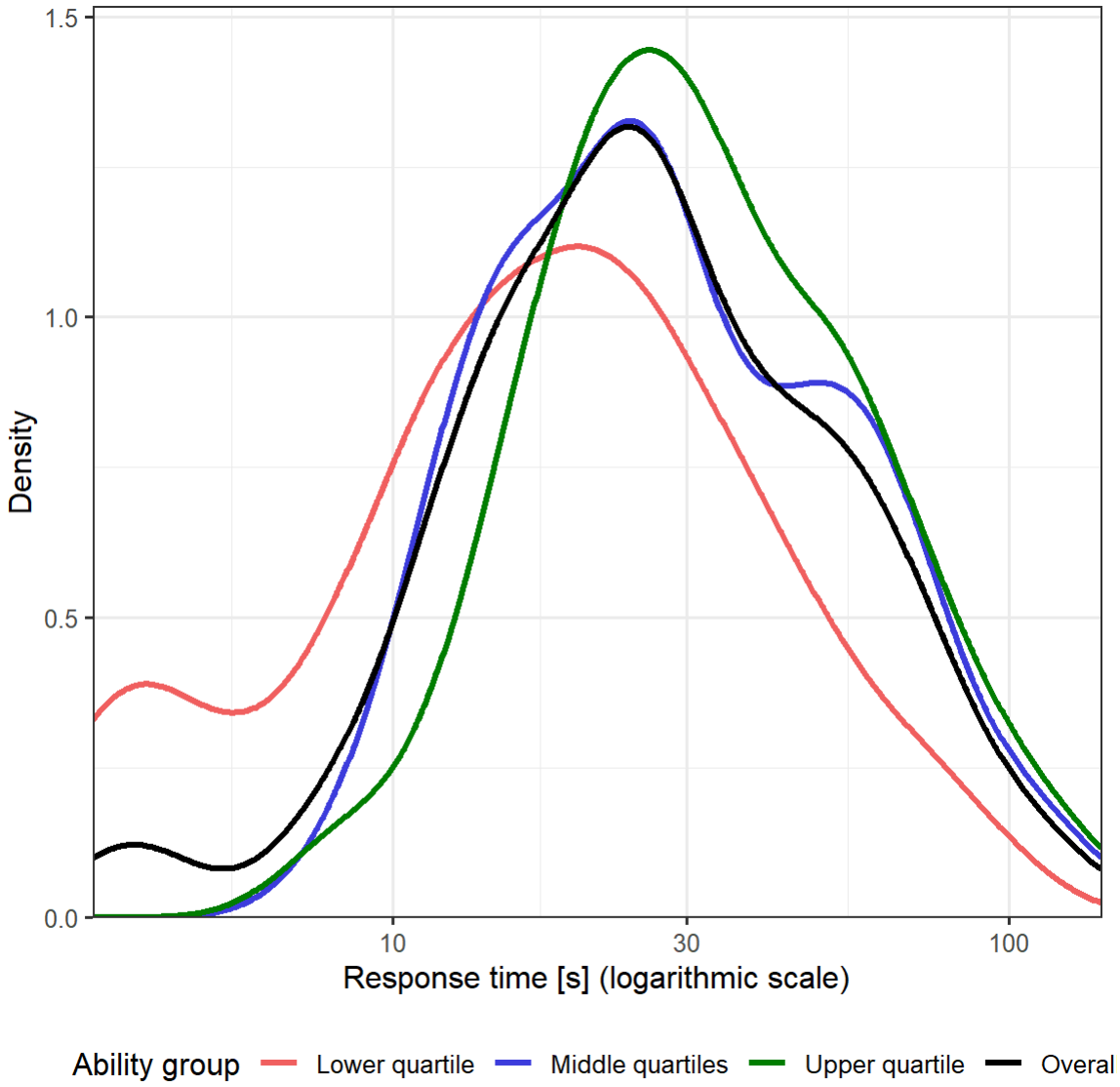
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
13	35	61	27	123	50.0%	51.3%	69.2%	53.9%
20	20	29	4	53	28.6%	24.4%	10.3%	23.2%
30	15	29	8	52	21.4%	24.4%	20.5%	22.8%
Overall	70	119	39	228	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



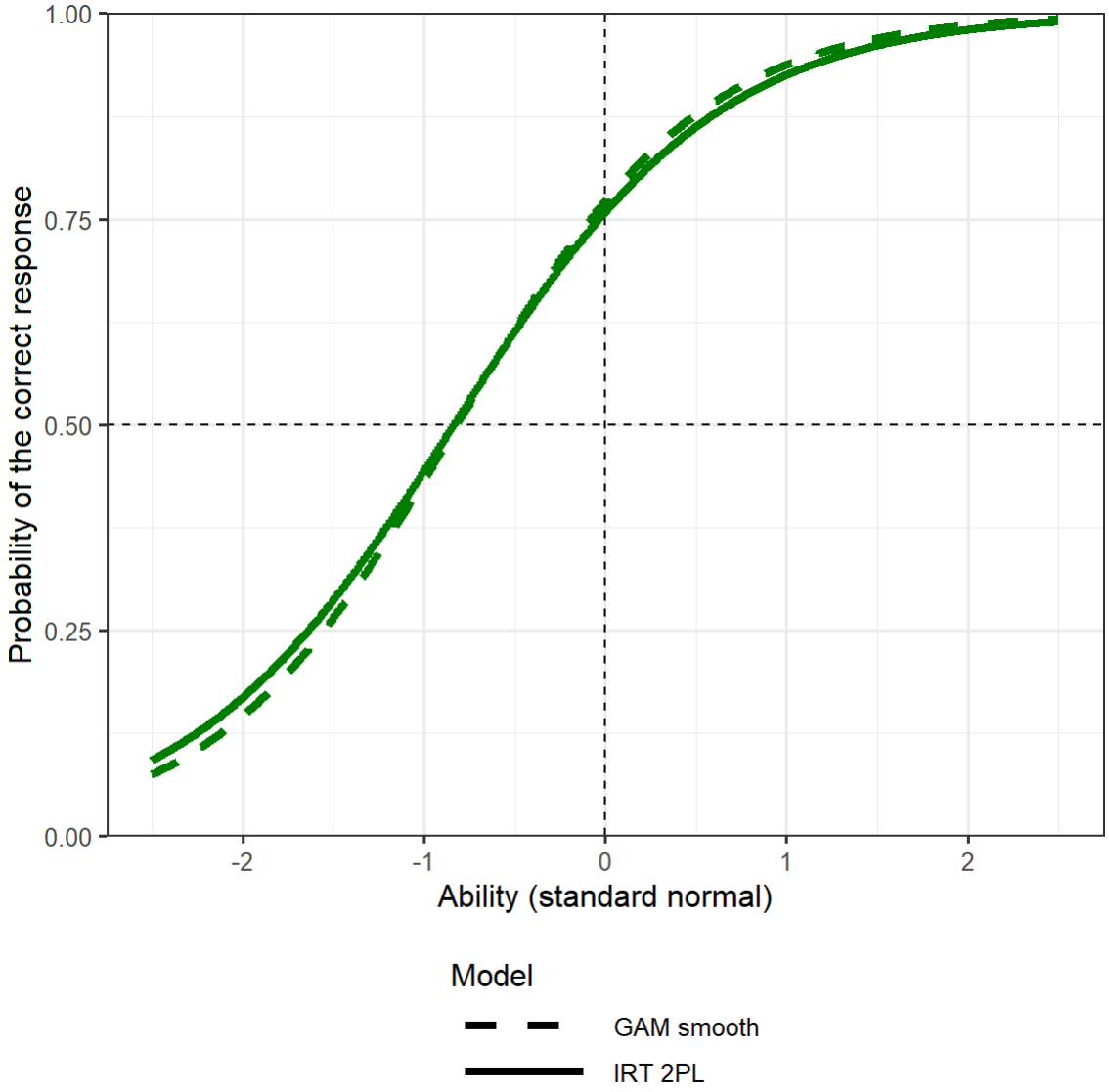
10.1.5.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.3	10.6	19.0	29.1	93.5	2.8	0.8
Middle quartiles	7.7	16.6	26.5	47.4	141.1	3.3	0.6
Upper quartile	7.8	21.2	29.8	50.2	129.8	3.4	0.6
Overall	3.3	16.1	24.9	42.6	141.1	3.2	0.7

10.1.6 MP71041 (TIMSS)

10.1.6.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	71.09%	294	TIMSS	MC	0.51	0.50	1.37	−0.84	0.000

10.1.6.2 Gender DIF:

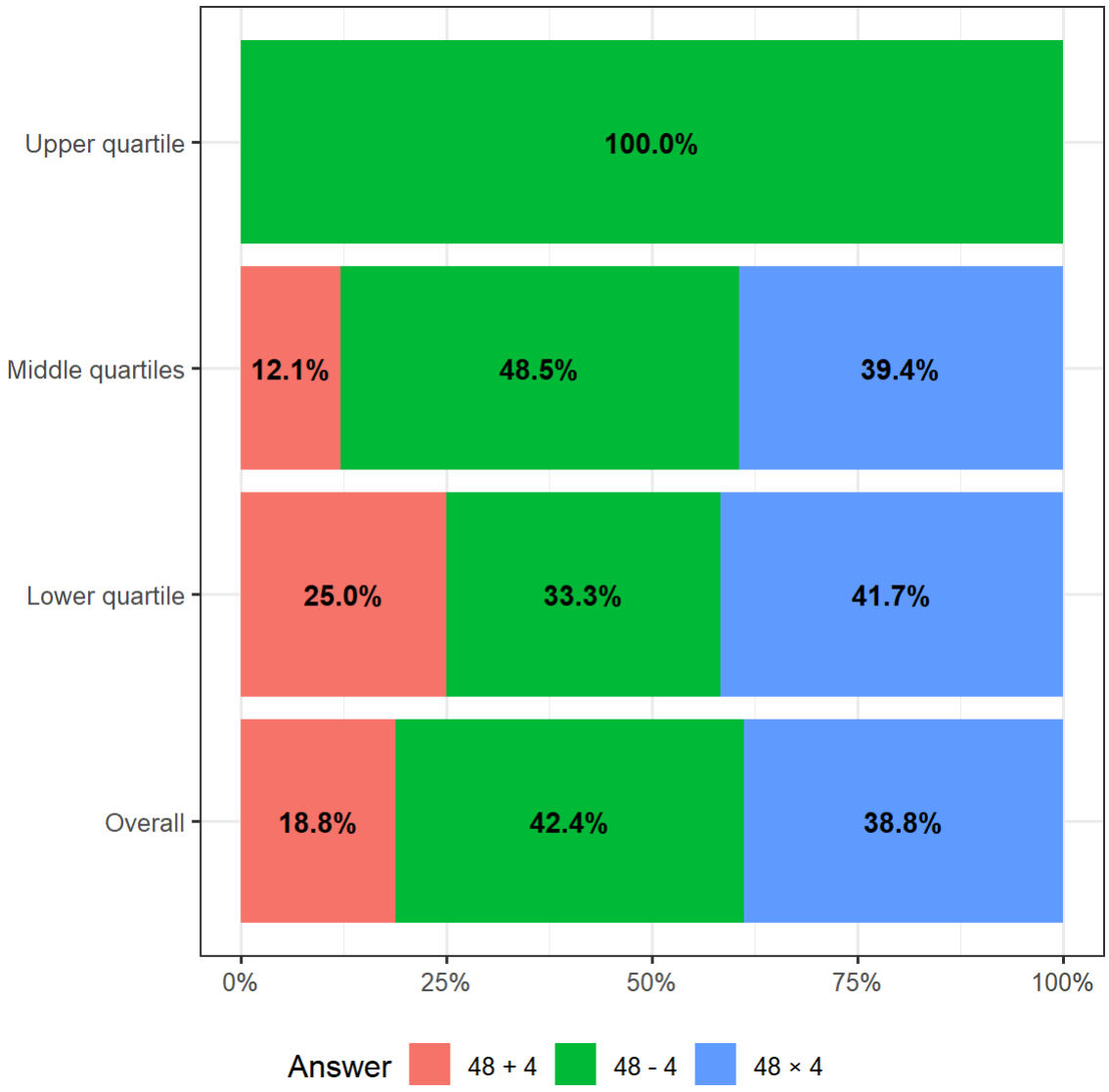
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.45	0.797	1.000	0.001	N/A

10.1.6.3 Frequency of answers:

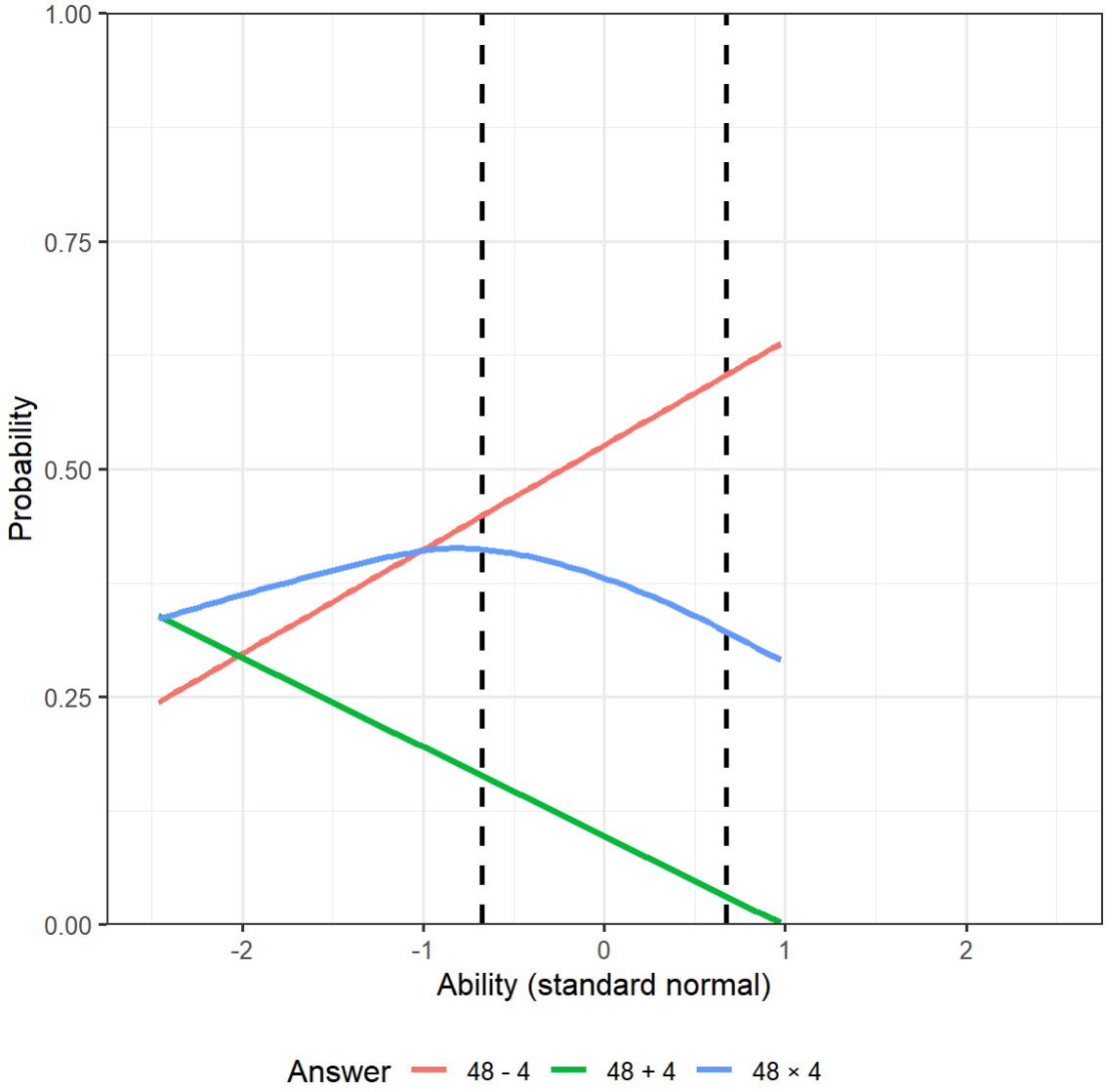
Answer	Frequency	Percent
48 + 4	16	5.4%
48 - 4	36	12.2%
48 × 4	33	11.2%
48 ÷ 4 (correct)	209	71.1%
Overall	294	100.0%

10.1.6.4 Distractors functioning:

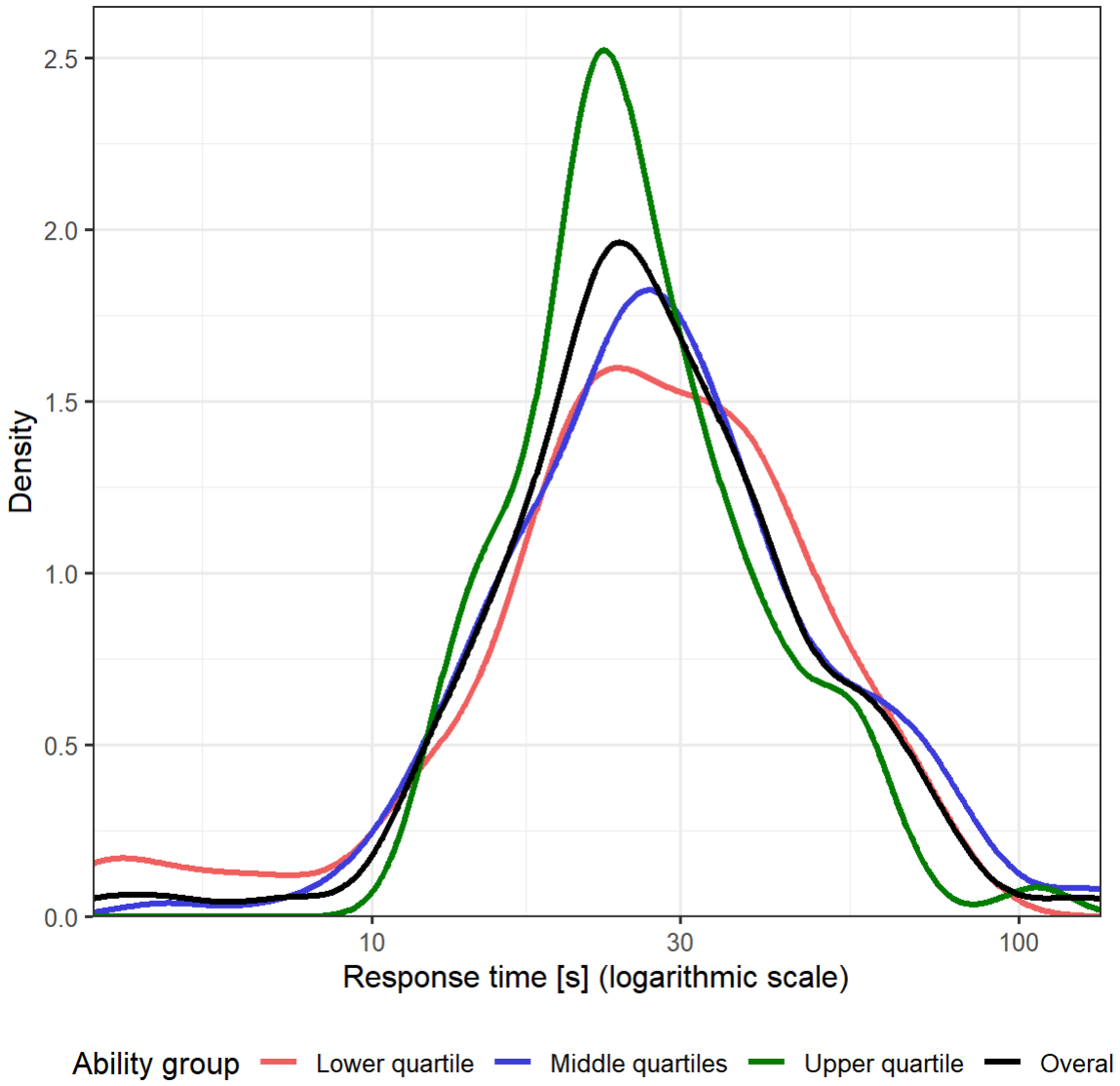
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
48 + 4	12	4	0	16	25.0%	12.1%	0.0%	18.8%
48 - 4	16	16	4	36	33.3%	48.5%	100.0%	42.4%
48 × 4	20	13	0	33	41.7%	39.4%	0.0%	38.8%
Overall	48	33	4	85	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



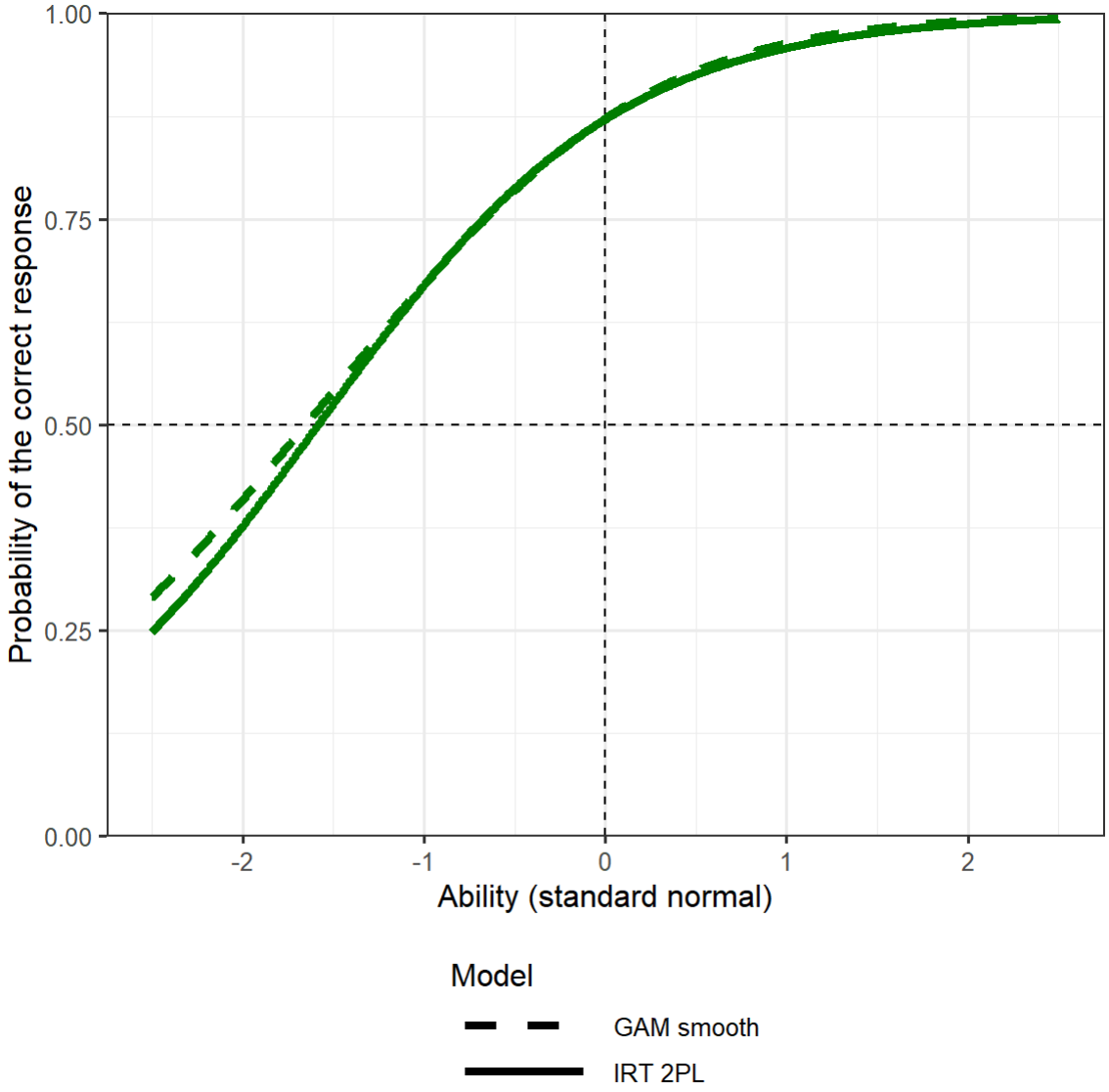
10.1.6.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.7	19.9	26.4	39.2	73.7	3.2	0.6
Middle quartiles	4.8	19.6	27.5	39.0	133.5	3.3	0.6
Upper quartile	12.5	20.1	24.1	32.3	106.6	3.3	0.4
Overall	3.7	19.9	26.0	37.2	133.5	3.3	0.5

10.1.7 M041191 (TIMSS)

10.1.7.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	82.65%	294	TIMSS	MC	0.40	0.41	1.21	−1.58	0.020

10.1.7.2 Gender DIF:

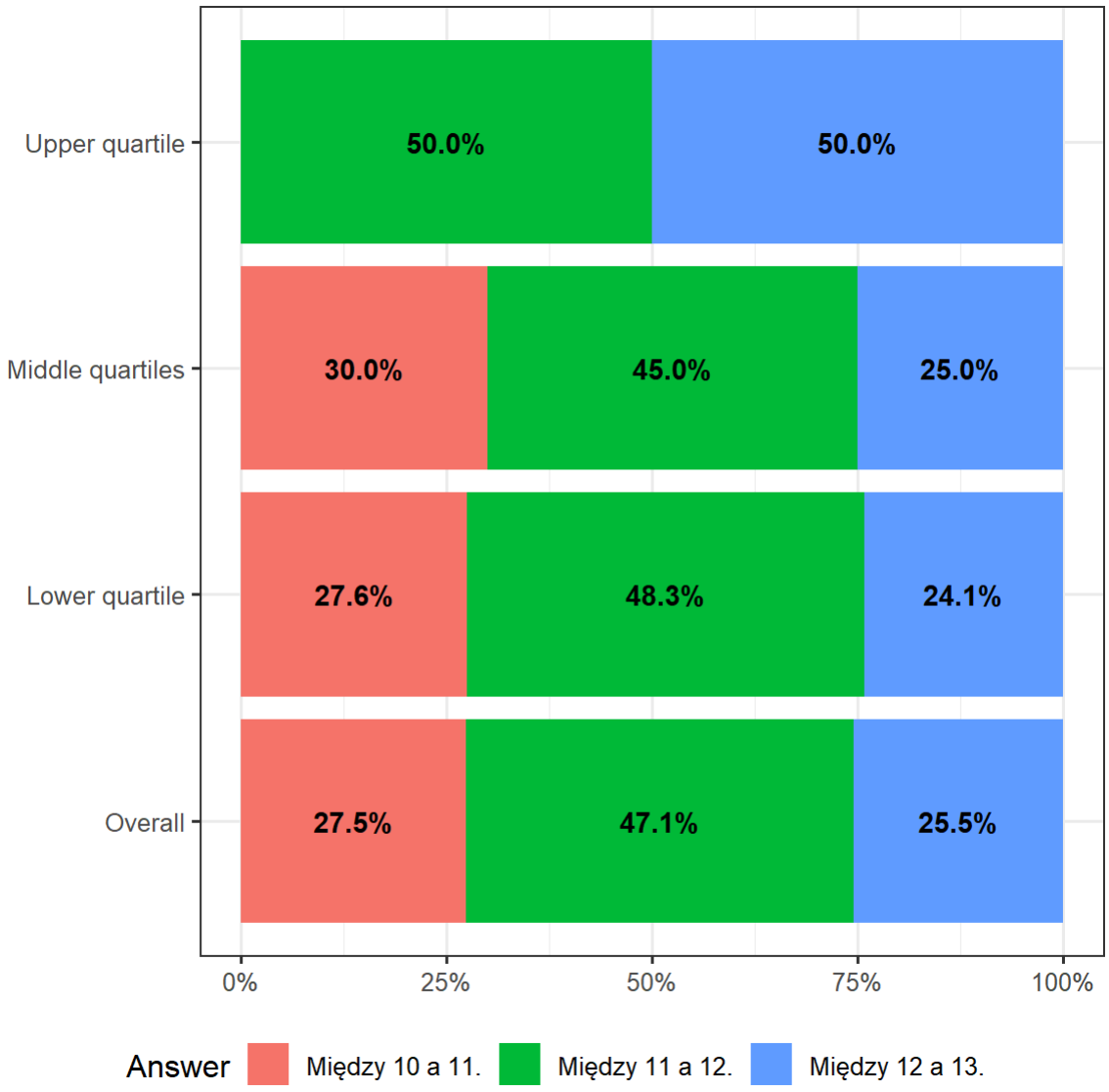
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.57	0.754	1.000	0.002	N/A

10.1.7.3 Frequency of answers:

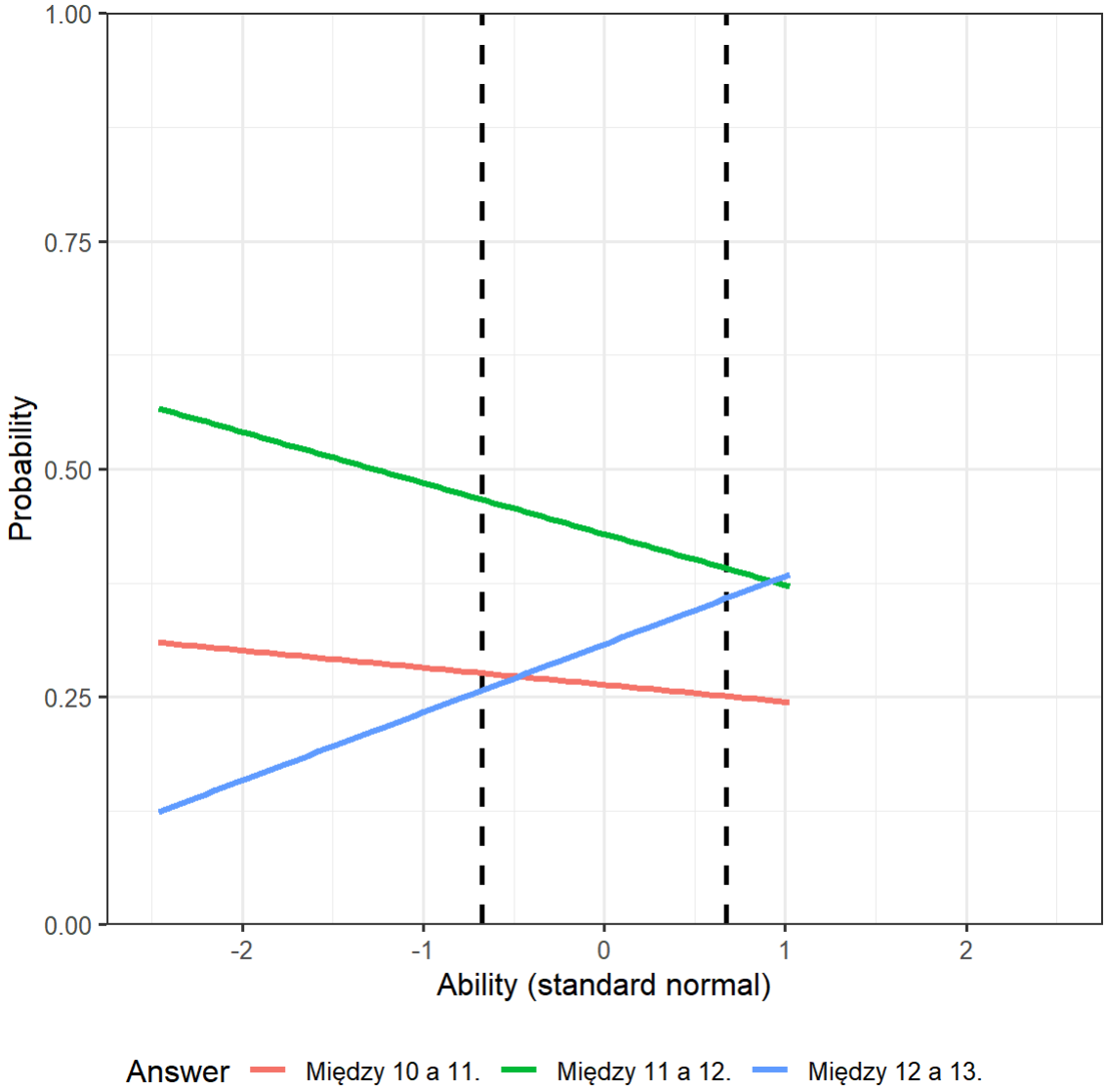
Answer	Frequency	Percent
Między 10 a 11.	14	4.8%
Między 11 a 12.	24	8.2%
Między 12 a 13.	13	4.4%
Między 13 a 14. (correct)	243	82.7%
Overall	294	100.0%

10.1.7.4 Distractors functioning:

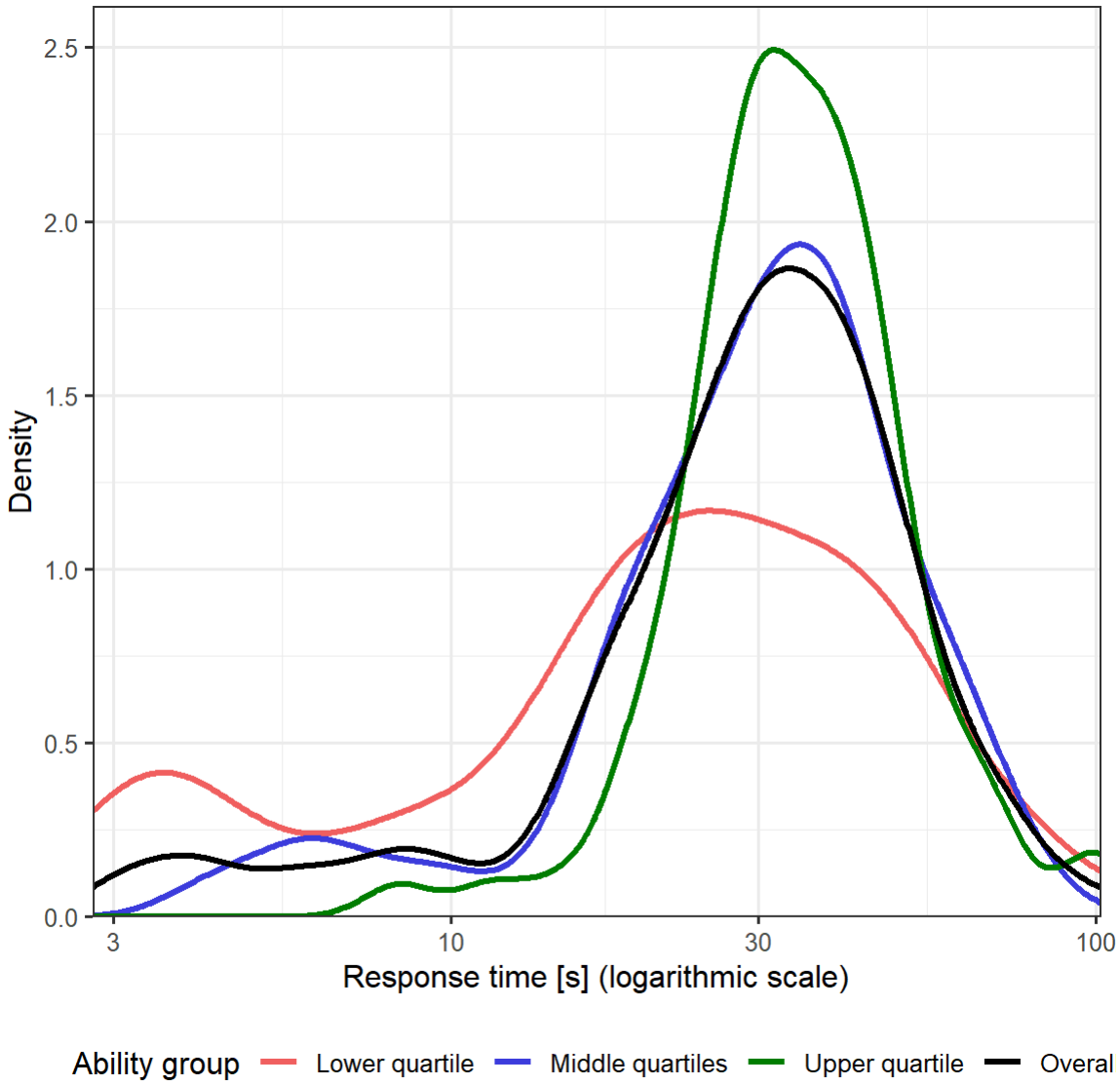
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Między 10 a 11.	8	6	0	14	27.6%	30.0%	0.0%	27.5%
Między 11 a 12.	14	9	1	24	48.3%	45.0%	50.0%	47.1%
Między 12 a 13.	7	5	1	13	24.1%	25.0%	50.0%	25.5%
Overall	29	20	2	51	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



10.1.7.5 Answering time:



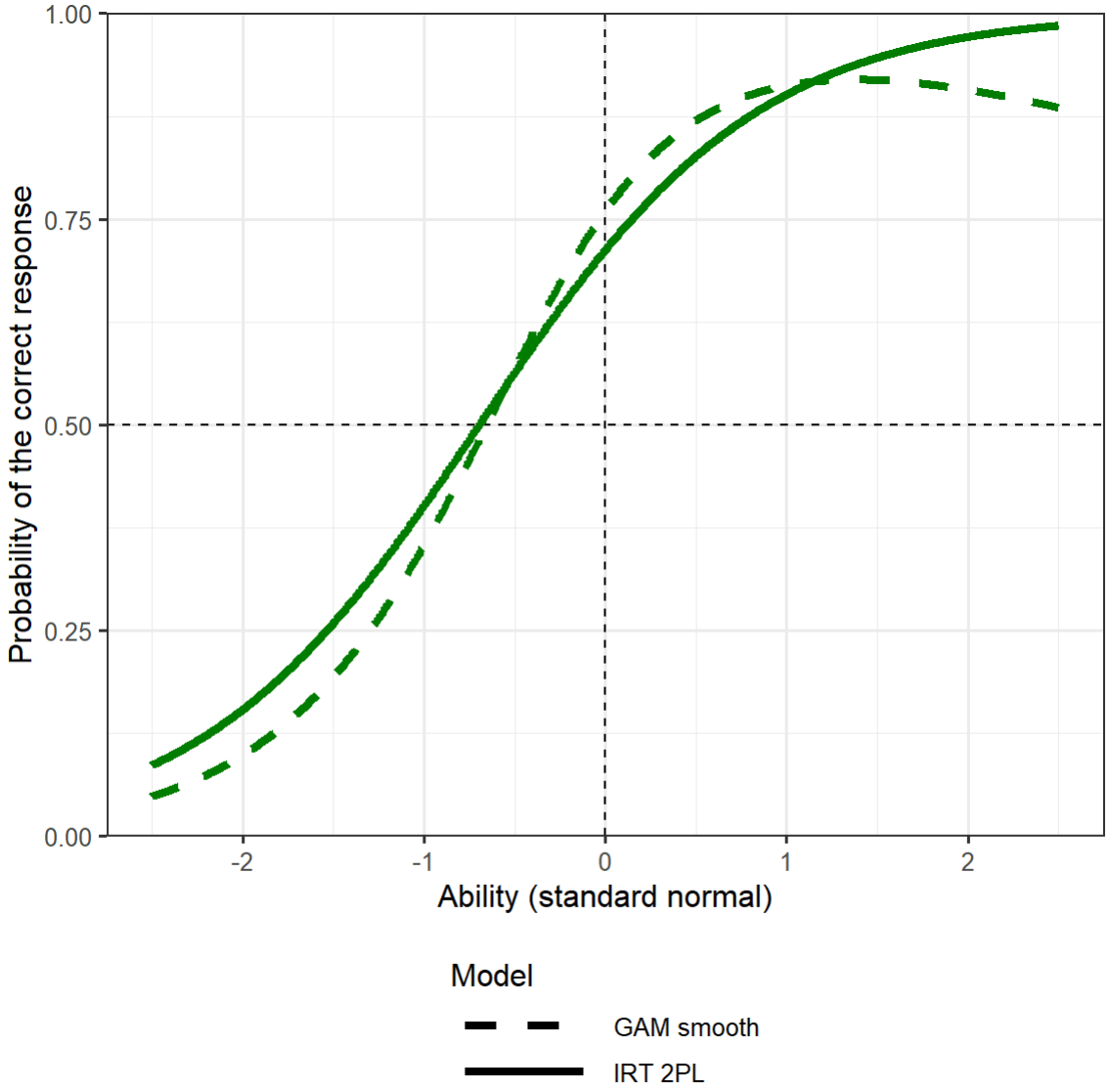
	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.8	14.4	23.7	37.1	89.7	3.0	0.9
Middle quartiles	4.0	22.4	31.2	41.7	86.5	3.4	0.6
Upper quartile	8.3	27.9	34.3	43.4	101.5	3.5	0.4
Overall	2.8	21.8	30.4	42.2	101.5	3.3	0.7

10.1.8 M061106 (TIMSS)

10.1.8.1 Item fit:

```
## Warning in newton(lsp = lsp, X = G$X, y = G$y, Eb = G$Eb, UrS = G$UrS, L = G$L,
## : Fitting terminated with step failure - check results carefully
```

```
## Warning in newton(lsp = lsp, X = G$X, y = G$y, Eb = G$Eb, UrS = G$UrS, L = G$L,
## : Fitting terminated with step failure - check results carefully
```



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	67.45%	299	TIMSS	MC	0.50	0.48	1.31	−0.69	0.018

10.1.8.2 Gender DIF:

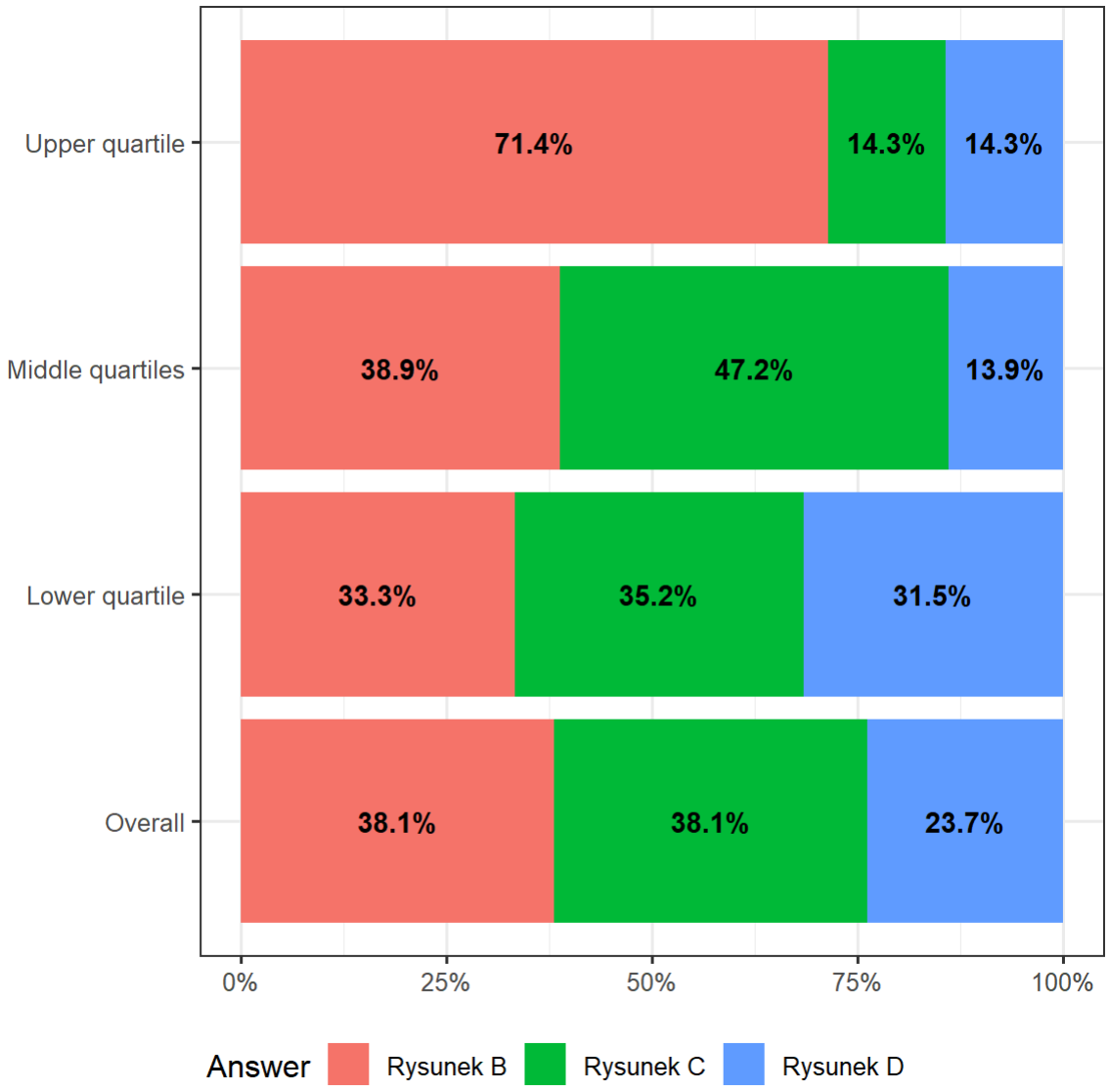
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
6.21	0.045	1.000	0.018	N/A

10.1.8.3 Frequency of answers:

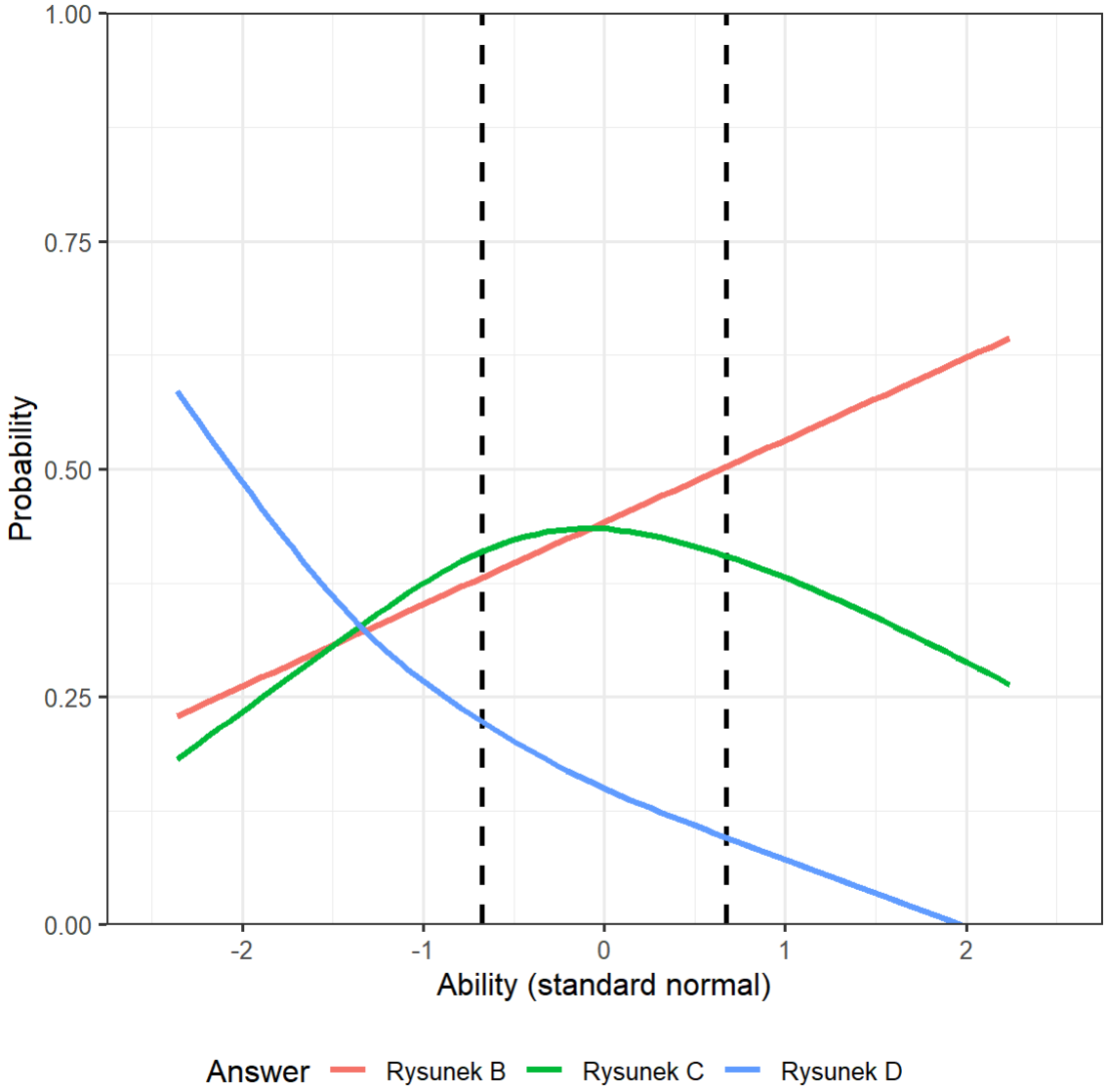
Answer	Frequency	Percent
Rysunek A (correct)	201	67.4%
Rysunek B	37	12.4%
Rysunek C	37	12.4%
Rysunek D	23	7.7%
Overall	298	100.0%

10.1.8.4 Distractors functioning:

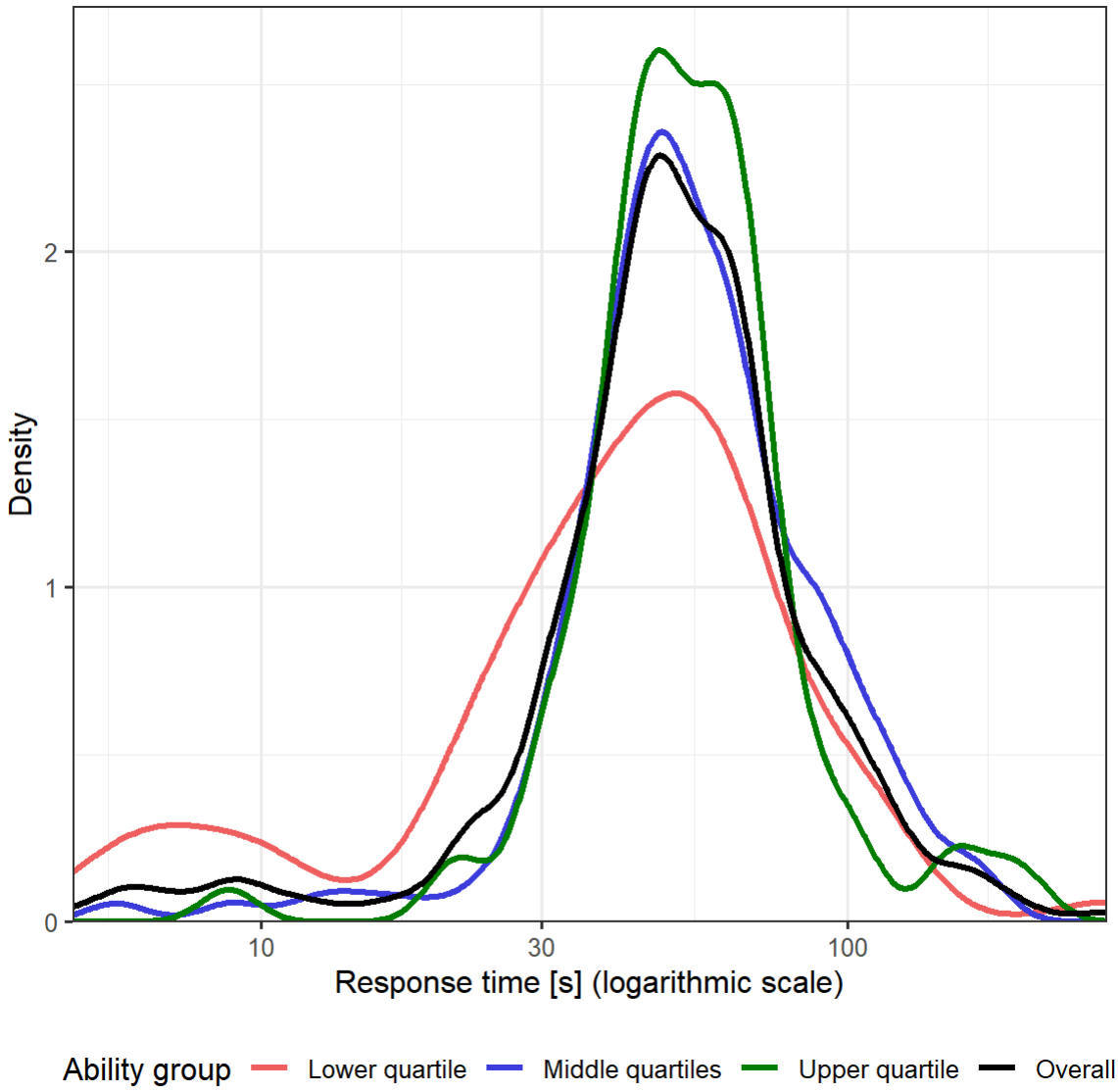
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Rysunek B	18	14	5	37	33.3%	38.9%	71.4%	38.1%
Rysunek C	19	17	1	37	35.2%	47.2%	14.3%	38.1%
Rysunek D	17	5	1	23	31.5%	13.9%	14.3%	23.7%
Overall	54	36	7	97	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



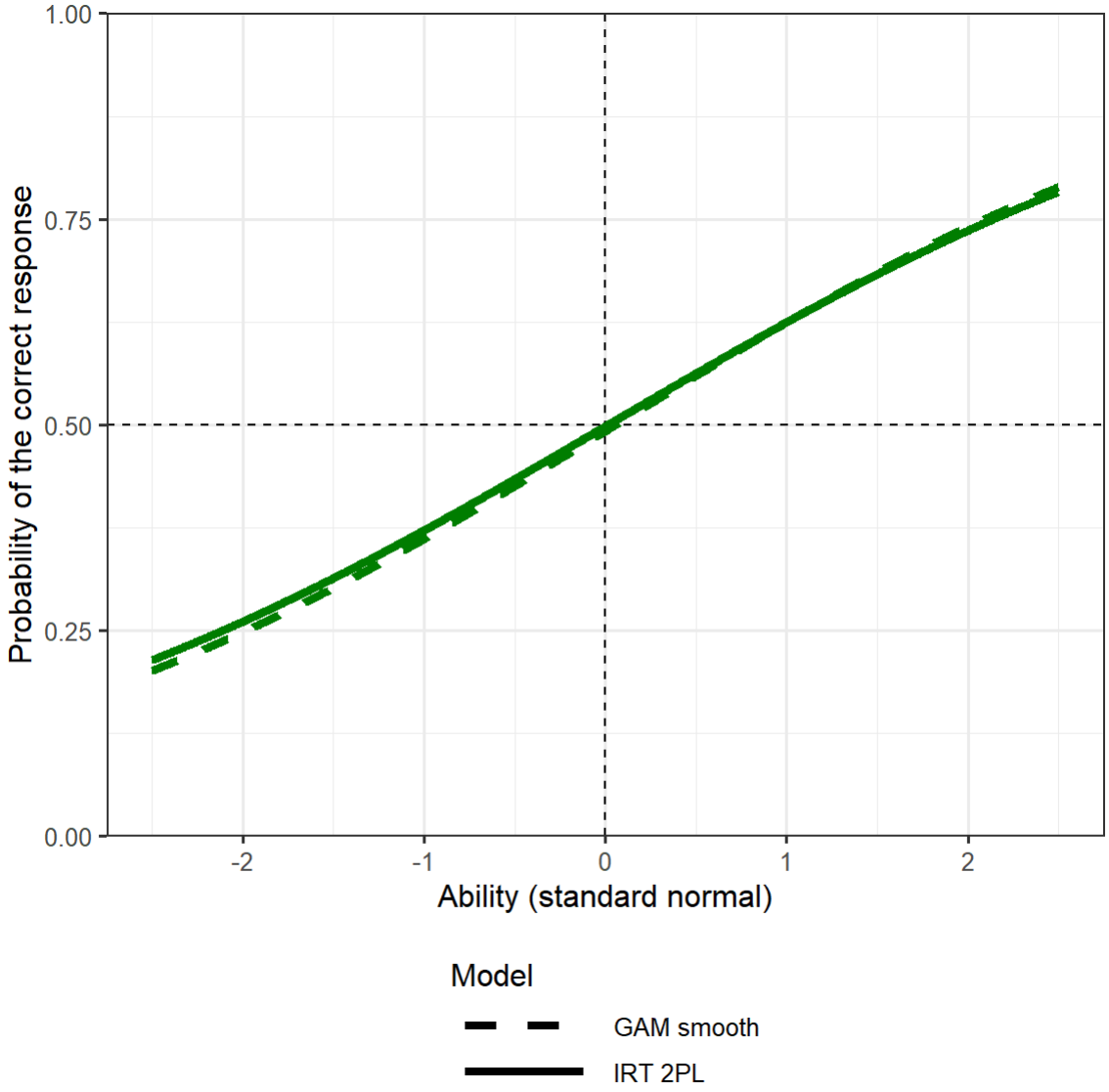
10.1.8.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.8	29.7	44.9	62.2	275.4	3.7	0.8
Middle quartiles	5.7	41.9	52.3	68.2	167.3	4.0	0.5
Upper quartile	8.8	42.7	51.9	63.8	203.1	4.0	0.4
Overall	4.8	39.3	50.3	66.1	275.4	3.9	0.6

10.1.9 M041280 (TIMSS)

10.1.9.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	49.34%	304	TIMSS	MC	0.26	0.24	0.52	0.02	0.000

10.1.9.2 Gender DIF:

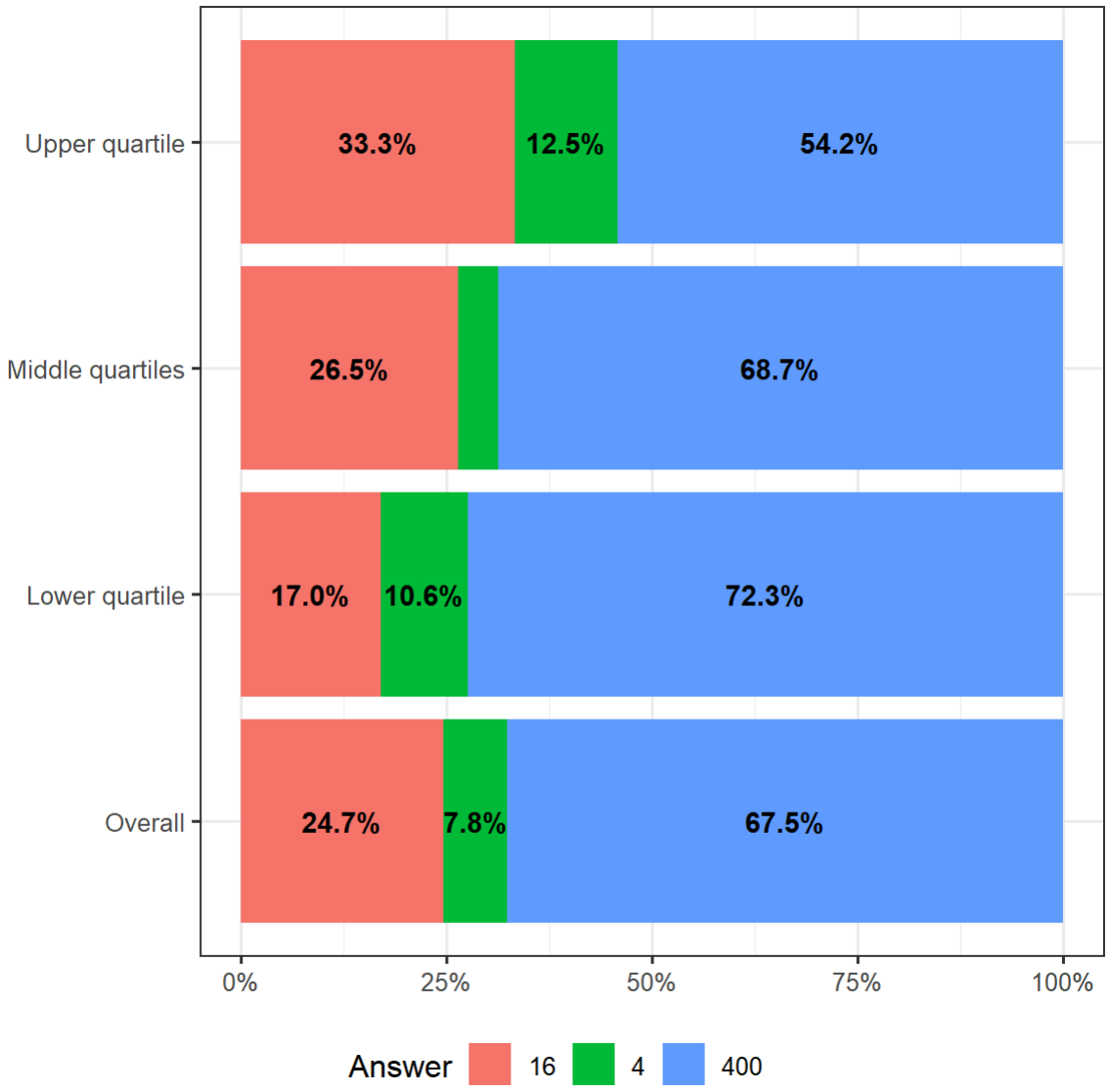
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.03	0.596	1.000	0.003	N/A

10.1.9.3 Frequency of answers:

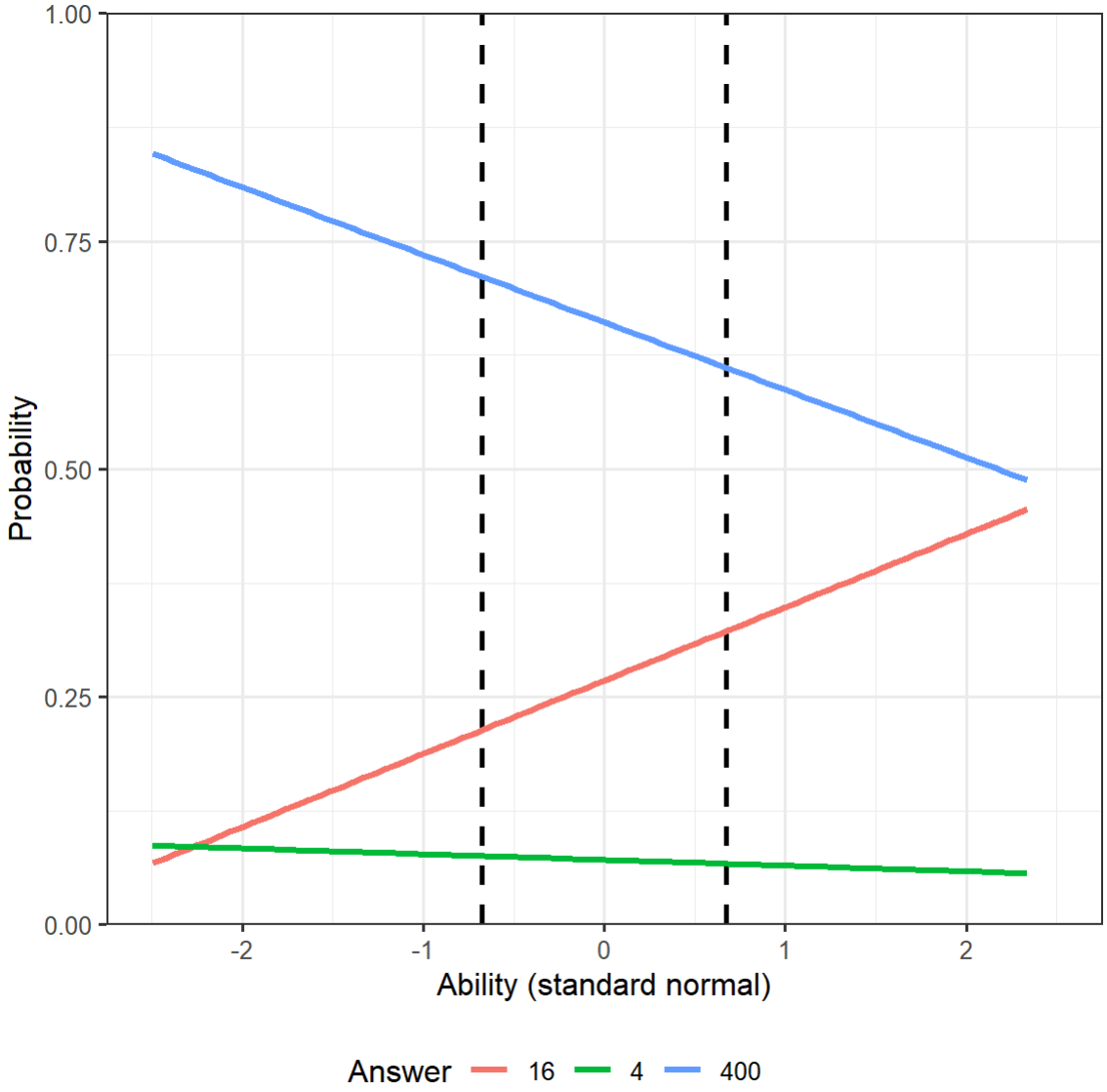
Answer	Frequency	Percent
4	12	3.9%
16	38	12.5%
40 (correct)	150	49.3%
400	104	34.2%
Overall	304	100.0%

10.1.9.4 Distractors functioning:

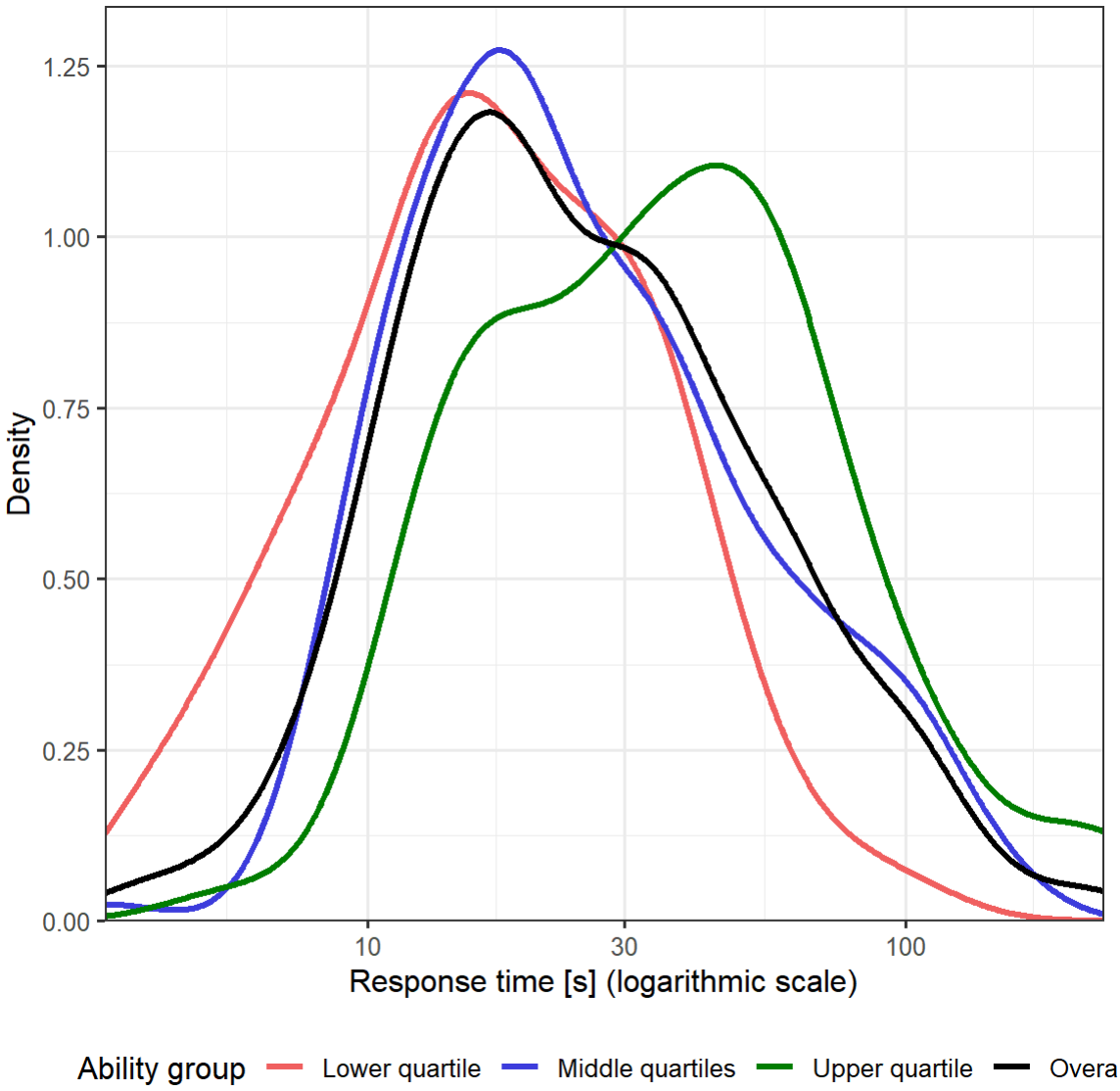
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
16	8	22	8	38	17.0%	26.5%	33.3%	24.7%
4	5	4	3	12	10.6%	4.8%	12.5%	7.8%
400	34	57	13	104	72.3%	68.7%	54.2%	67.5%
Overall	47	83	24	154	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



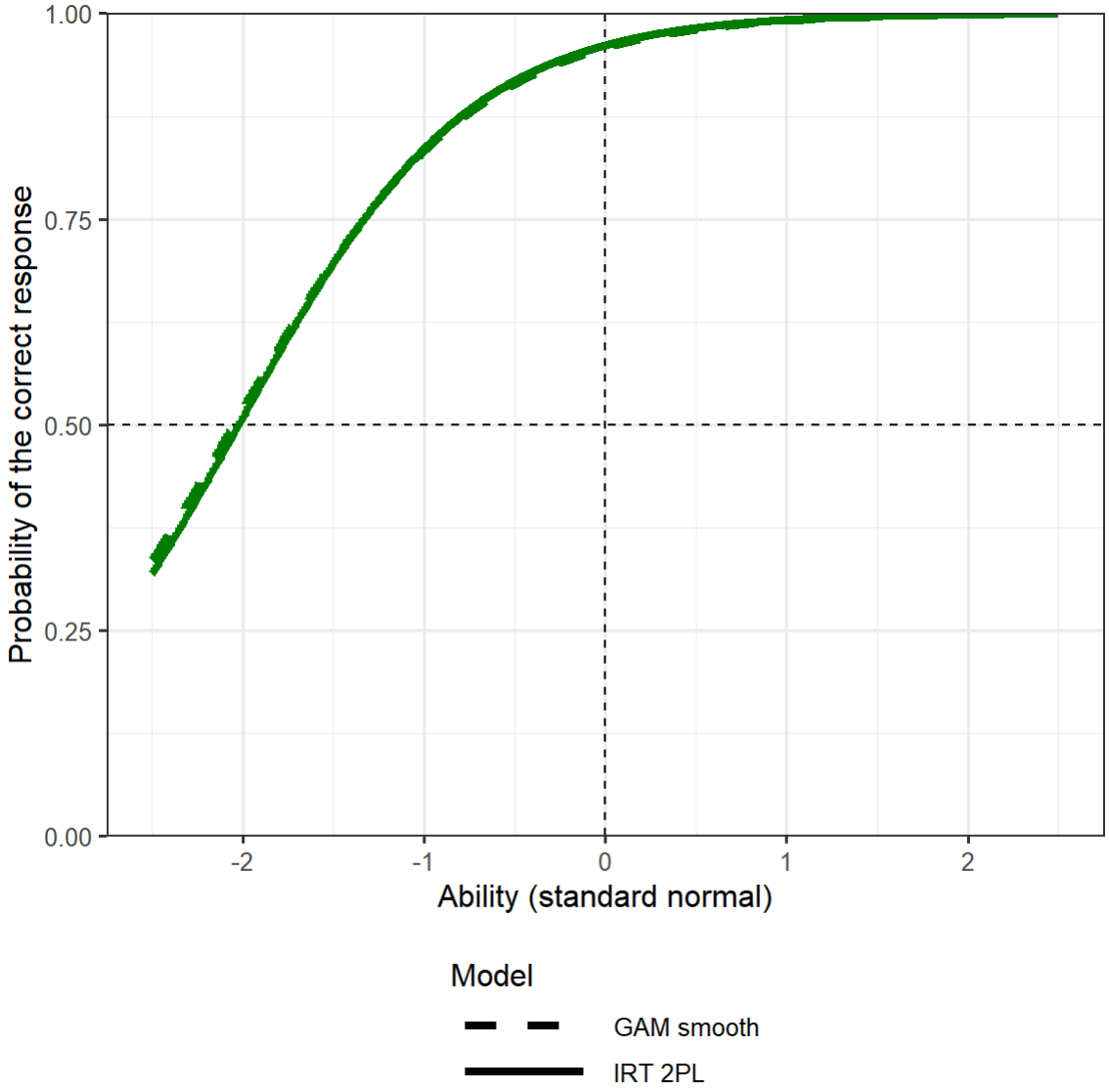
10.1.9.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.9	10.3	16.7	29.0	98.0	2.8	0.7
Middle quartiles	3.2	14.2	20.7	39.0	157.0	3.2	0.7
Upper quartile	5.6	18.8	36.4	57.4	233.4	3.6	0.8
Overall	3.2	14.1	22.2	41.3	233.4	3.2	0.8

10.1.10 M090055 (LLM)

10.1.10.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	91.45%	304	LLM teachers' selection	MC	0.38	0.25	1.59	−2.02	0.0(

10.1.10.2 Gender DIF:

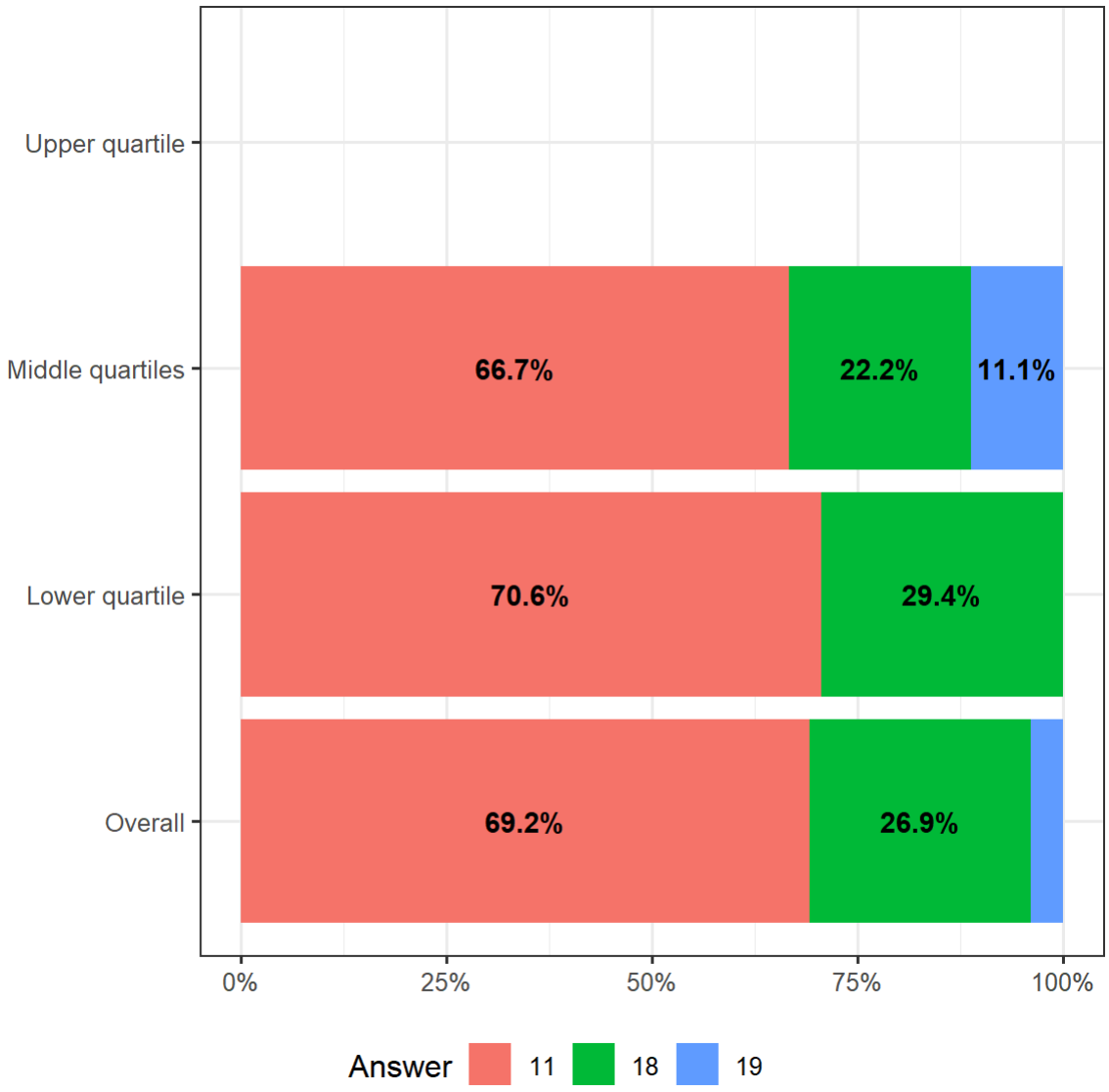
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.34	0.845	1.000	0.002	N/A

10.1.10.3 Frequency of answers:

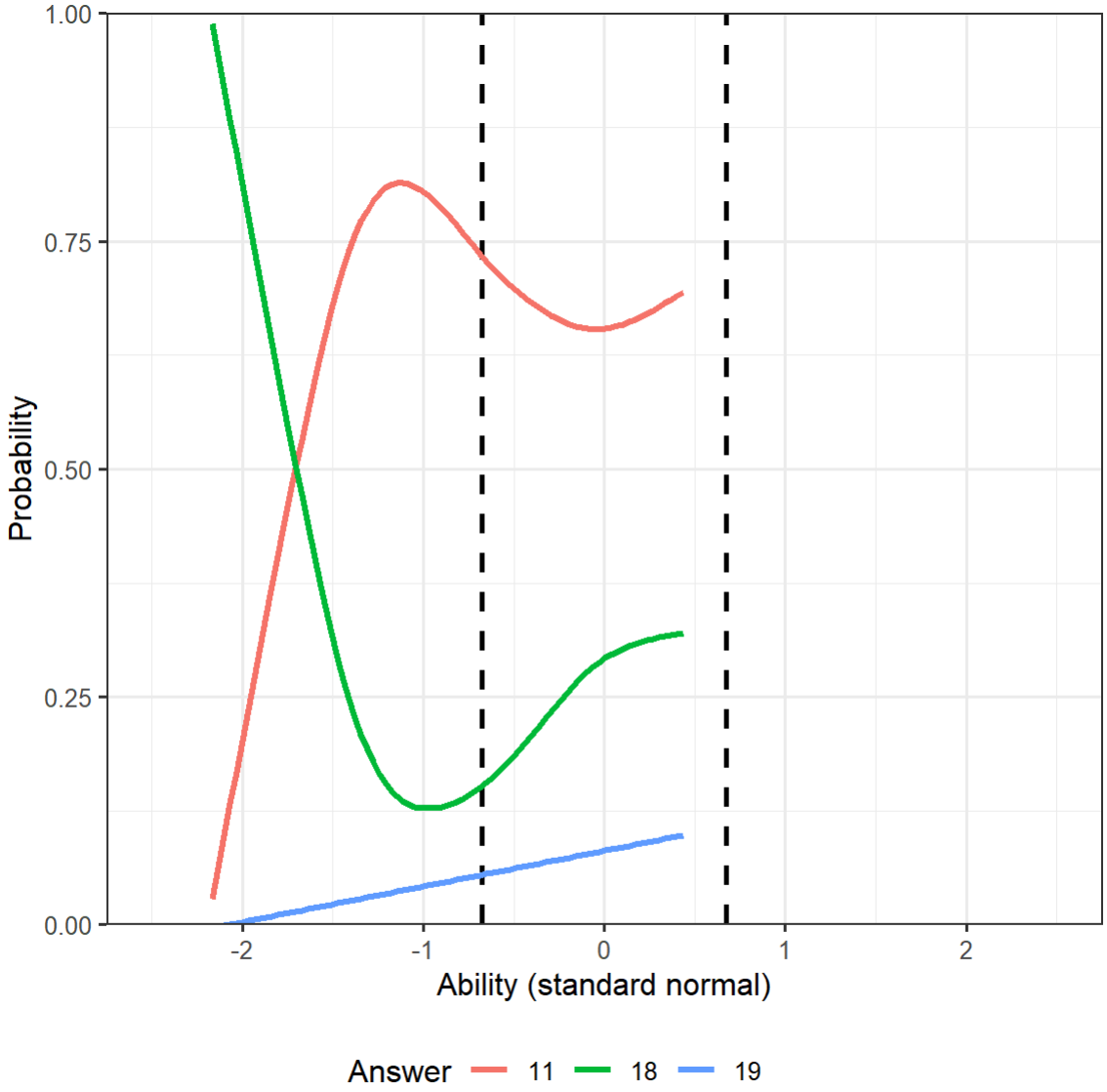
Answer	Frequency	Percent
11	18	5.9%
18	7	2.3%
19	1	0.3%
24 (correct)	278	91.4%
Overall	304	100.0%

10.1.10.4 Distractors functioning:

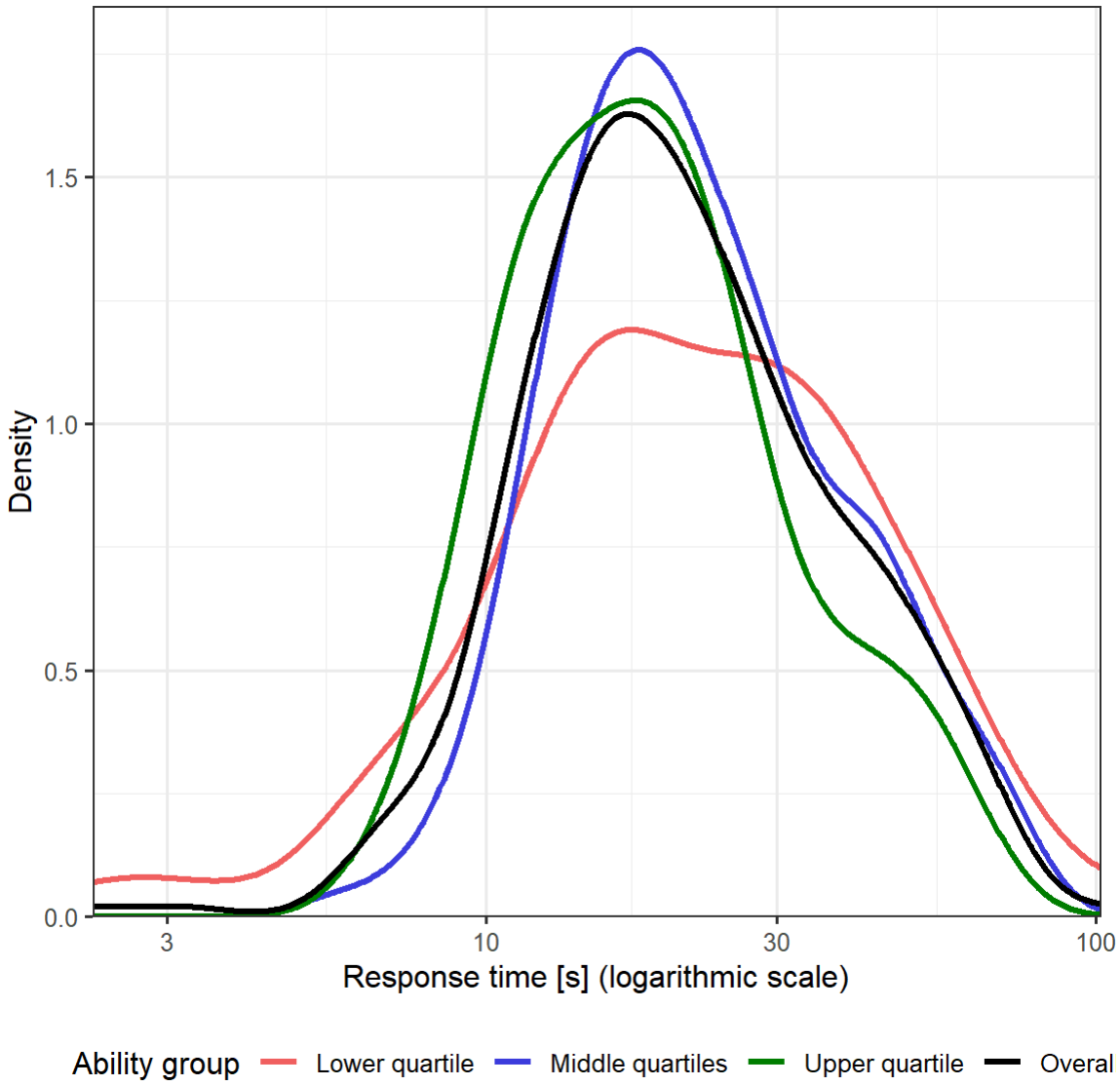
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
11	12	6	0	18	70.6%	66.7%	0.0%	69.2%
18	5	2	0	7	29.4%	22.2%	0.0%	26.9%
Overall	17	9	0	26	100.0%	100.0%	0.0%	100.0%
19	0	1	0	1	0.0%	11.1%	0.0%	3.8%




```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



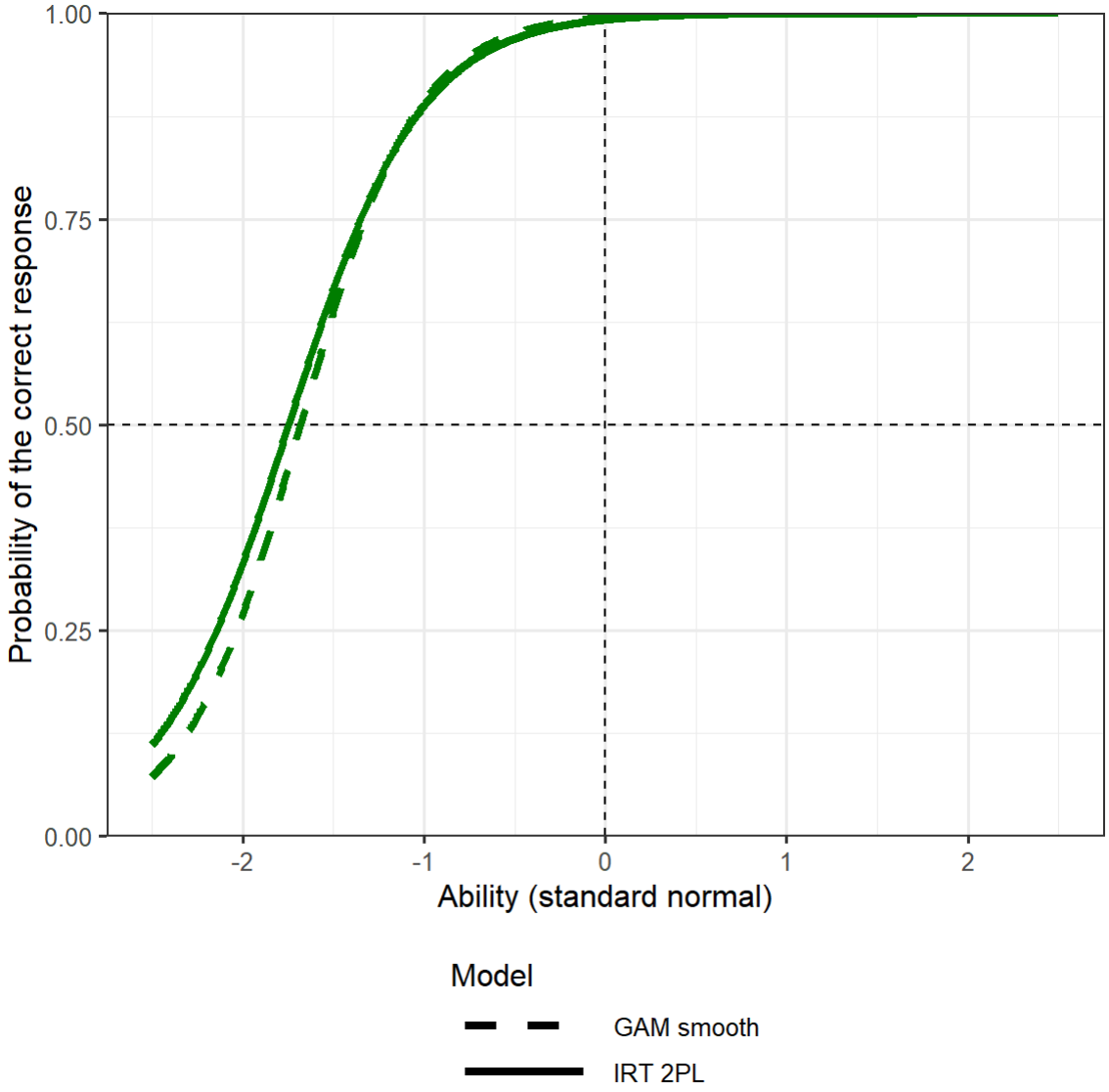
10.1.10.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.3	13.6	21.5	34.8	101.6	3.0	0.7
Middle quartiles	5.8	15.2	20.7	31.0	71.5	3.1	0.5
Upper quartile	6.8	12.5	18.2	25.0	61.3	2.9	0.5
Overall	2.3	14.1	19.9	31.1	101.6	3.0	0.6

10.1.11 M090006 (LLM)

10.1.11.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS generated items: Discrimination	IRT model including both TIMSS generated items: Item difficulty	IRT model including both TIMSS generated items: RMSE
1	93.09%	304	LLM teachers' selection	MC	0.48	0.29	2.77	−1.75	0.0(

10.1.11.2 Gender DIF:

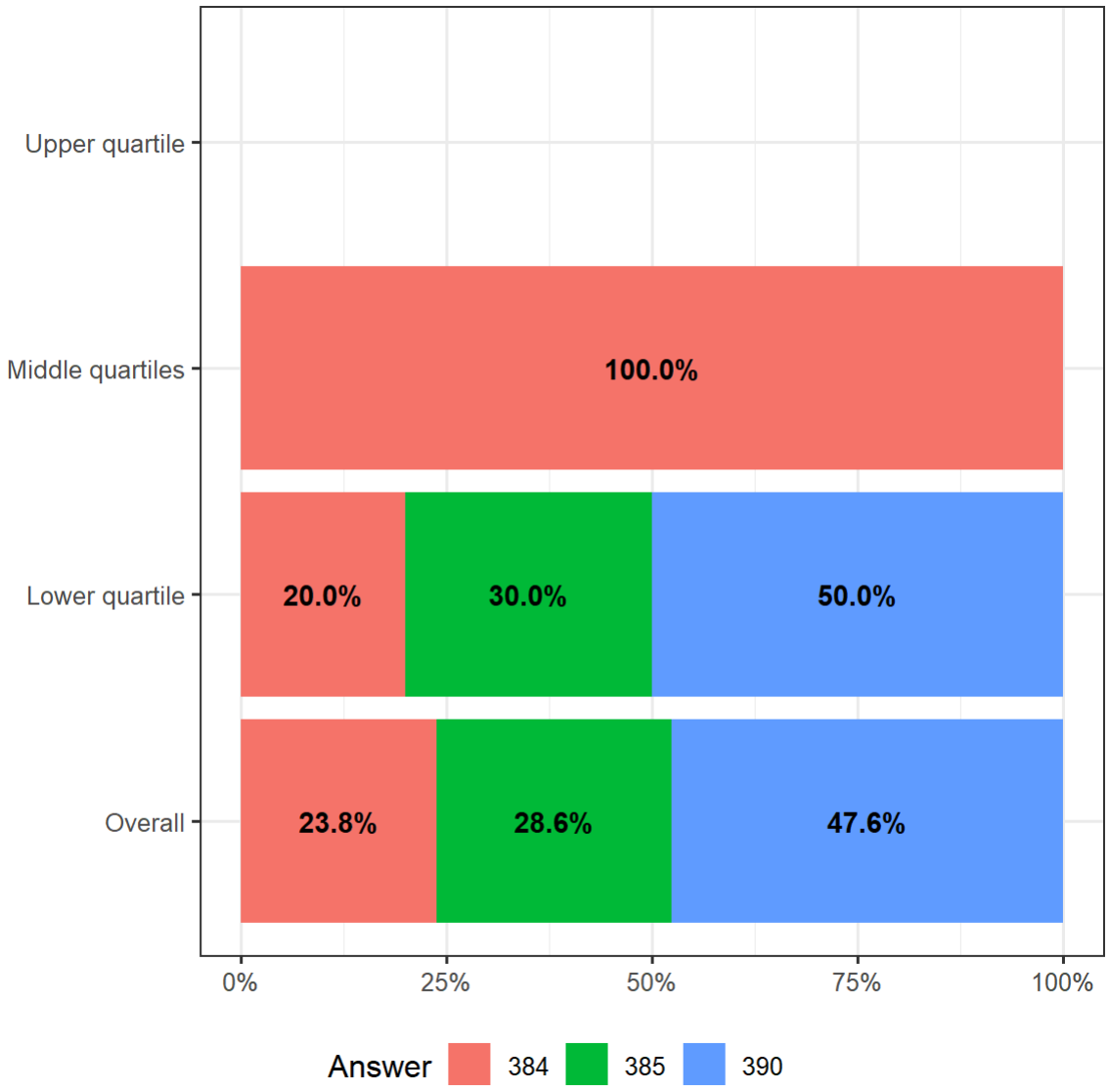
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.13	0.937	1.000	0.001	N/A

10.1.11.3 Frequency of answers:

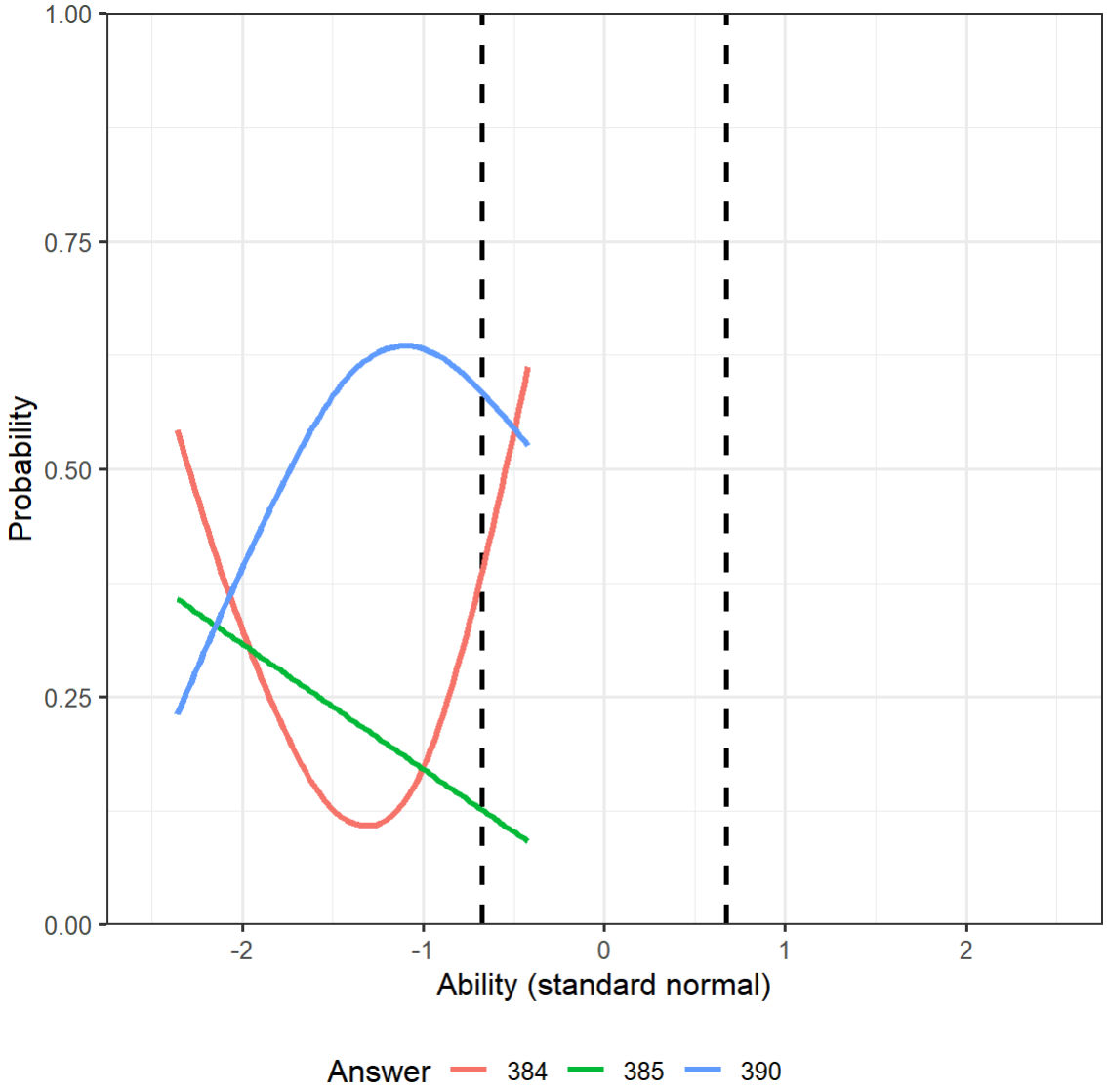
Answer	Frequency	Percent
384	5	1.6%
385	6	2.0%
386 (correct)	283	93.1%
390	10	3.3%
Overall	304	100.0%

10.1.11.4 Distractors functioning:

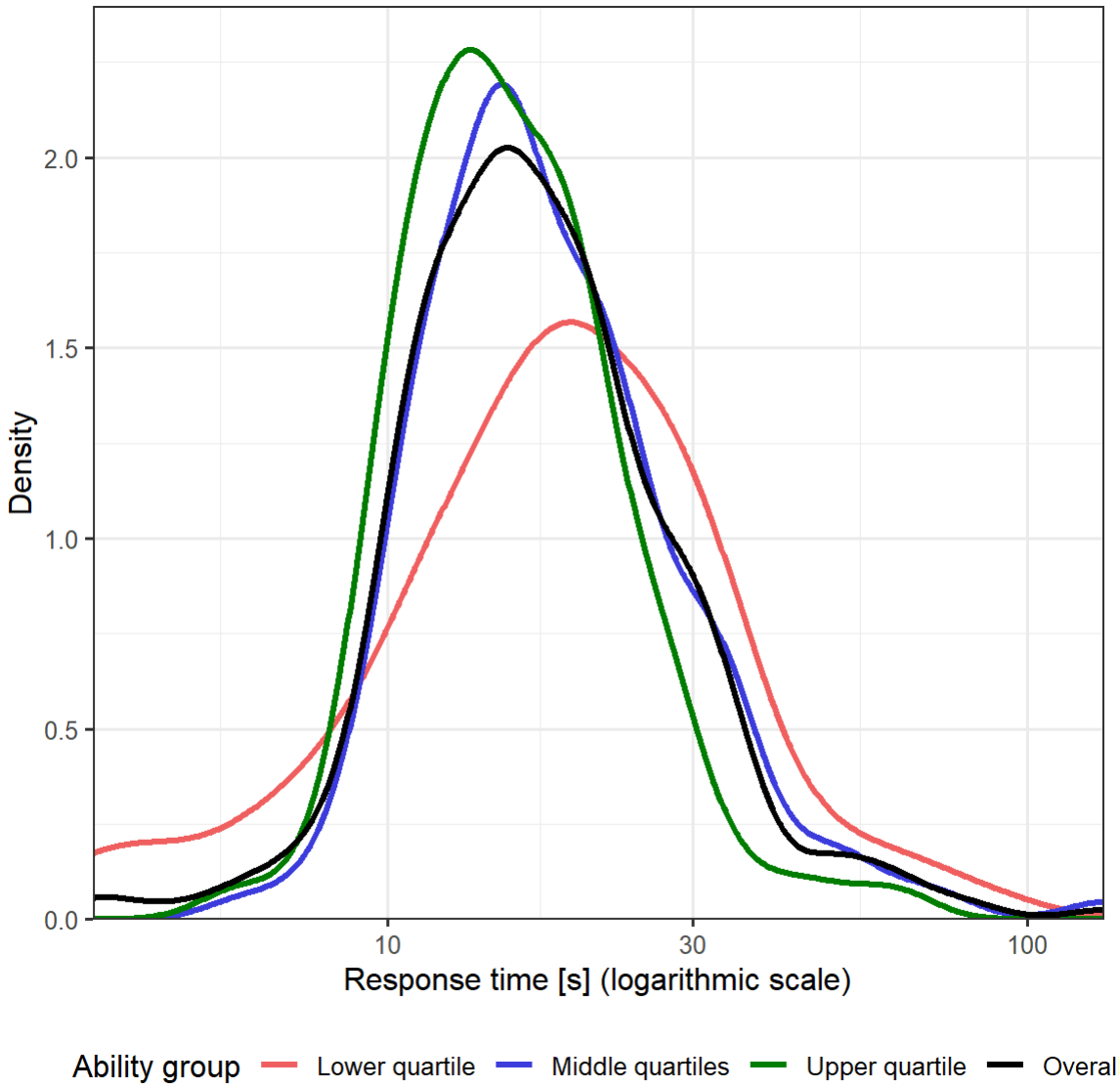
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
384	4	1	0	5	20.0%	100.0%	0.0%	23.8%
385	6	0	0	6	30.0%	0.0%	0.0%	28.6%
390	10	0	0	10	50.0%	0.0%	0.0%	47.6%
Overall	20	1	0	21	100.0%	100.0%	0.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



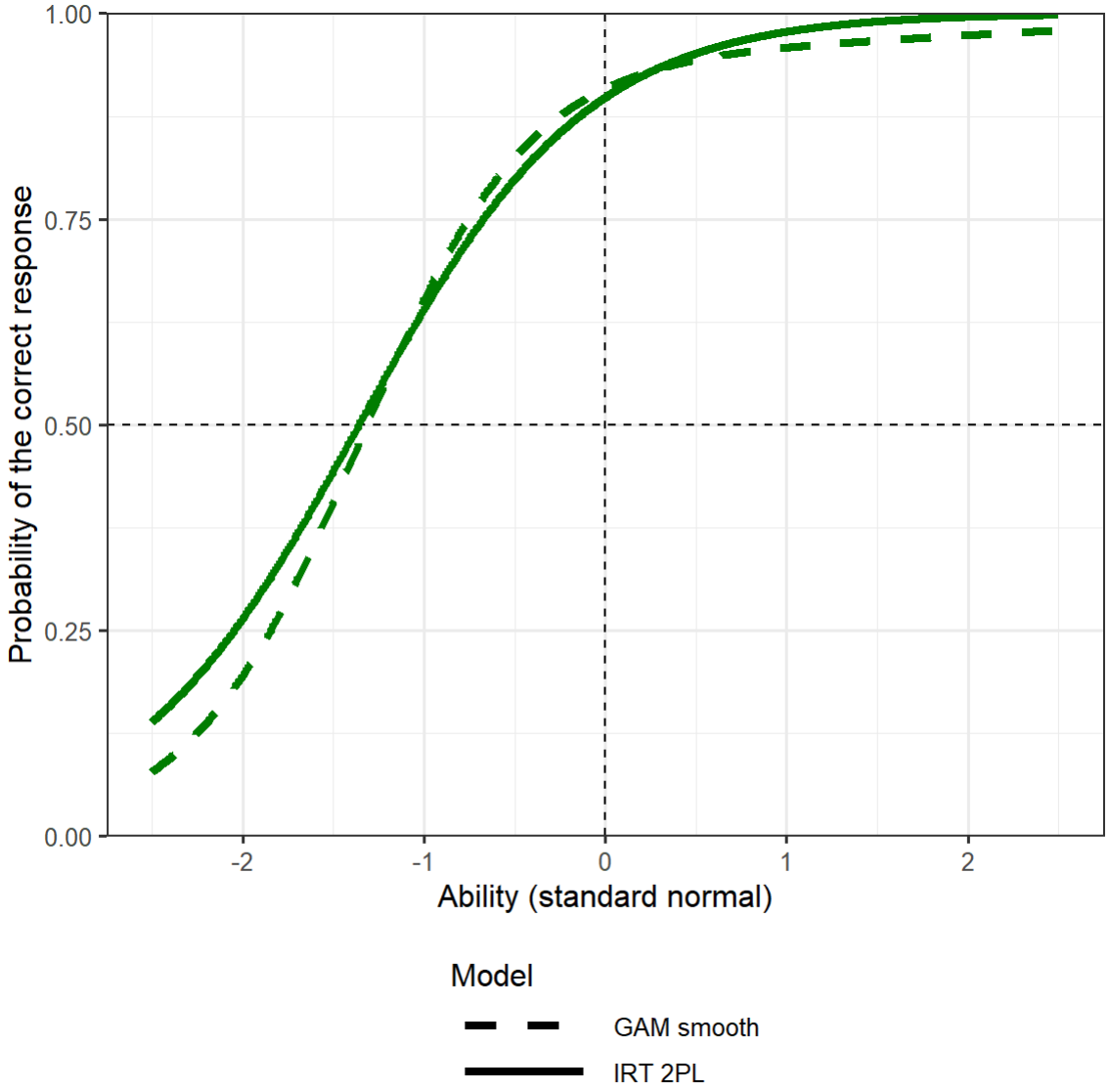
10.1.11.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	12.3	18.4	27.6	83.5	2.9	0.6
Middle quartiles	5.9	13.4	16.7	23.3	131.4	2.9	0.5
Upper quartile	5.8	11.7	15.0	19.6	61.9	2.7	0.4
Overall	3.5	12.4	16.5	22.3	131.4	2.8	0.5

10.1.12 M090059 (LLM)

10.1.12.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	82.89%	304	LLM teachers' selection	MC	0.46	0.28	1.60	−1.36	0.0(

10.1.12.2 Gender DIF:

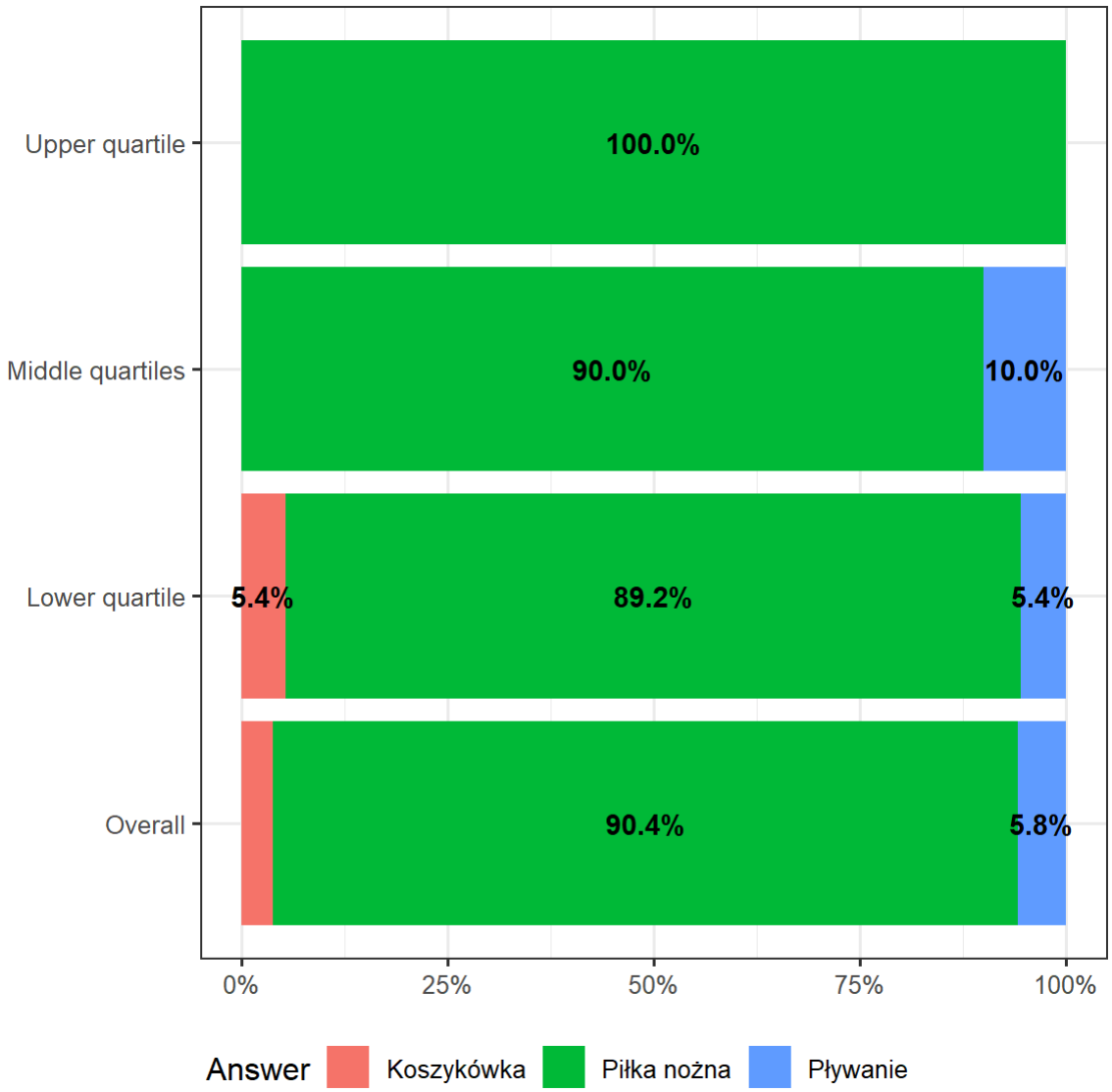
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.85	0.655	1.000	0.003	N/A

10.1.12.3 Frequency of answers:

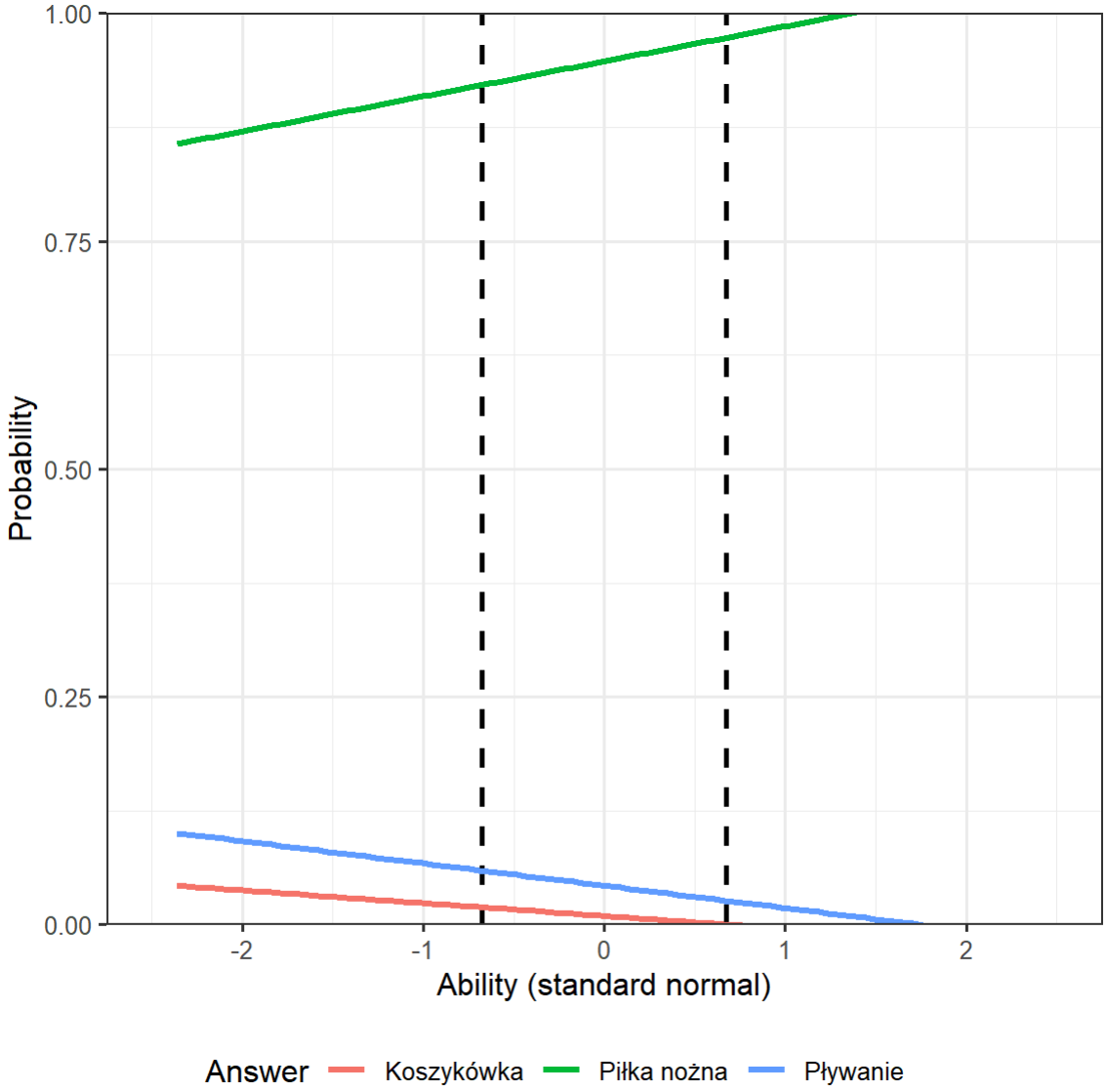
Answer	Frequency	Percent
Koszykówka	2	0.7%
Piłka nożna	47	15.5%
Pływanie	3	1.0%
Siatkówka (correct)	252	82.9%
Overall	304	100.0%

10.1.12.4 Distractors functioning:

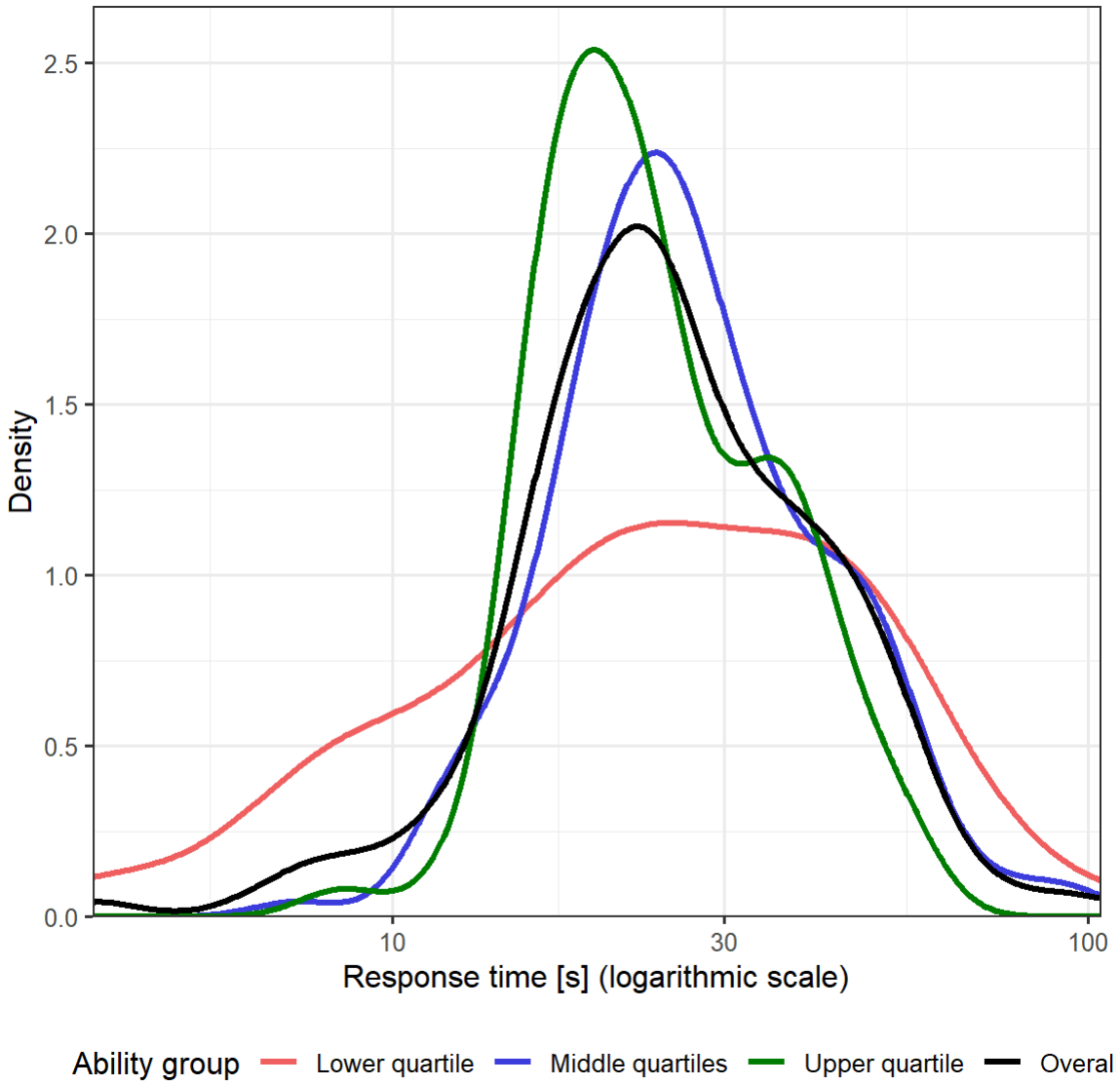
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Koszykówka	2	0	0	2	5.4%	0.0%	0.0%	3.8%
Piłka nożna	33	9	5	47	89.2%	90.0%	100.0%	90.4%
Pływanie	2	1	0	3	5.4%	10.0%	0.0%	5.8%
Overall	37	10	5	52	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



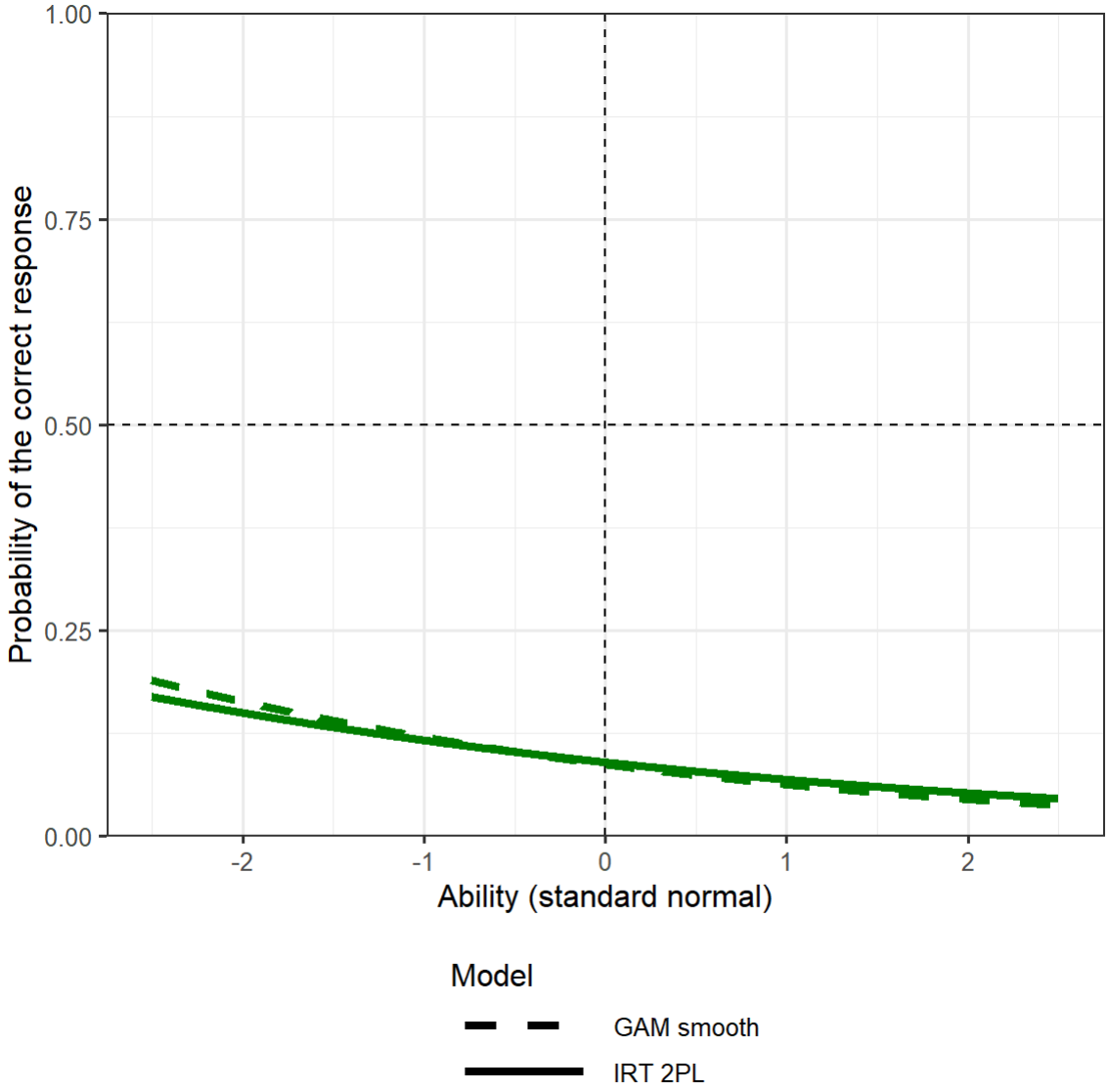
10.1.12.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.7	15.3	23.7	40.9	104.1	3.1	0.7
Middle quartiles	7.3	19.6	25.6	34.8	96.3	3.3	0.4
Upper quartile	8.4	18.0	22.2	31.9	55.8	3.2	0.4
Overall	3.7	18.3	24.1	35.2	104.1	3.2	0.5

10.1.13 M090017 (LLM)

10.1.13.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSI
1	9.29%	281	LLM teachers' selection	MC	−0.11	−0.05	−0.29	−7.97	0.0

10.1.13.2 Gender DIF:

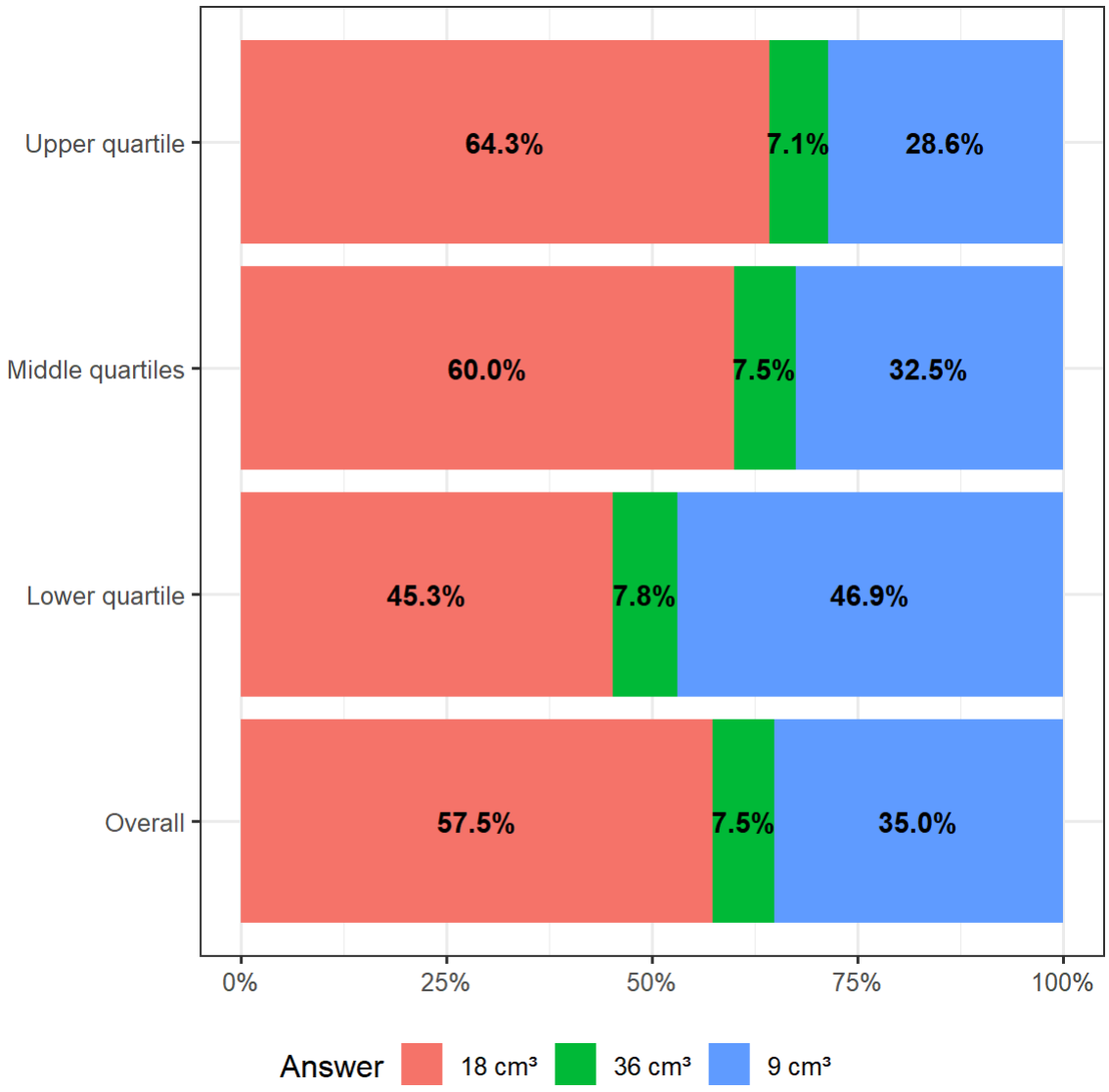
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.62	0.733	1.000	0.004	N/A

10.1.13.3 Frequency of answers:

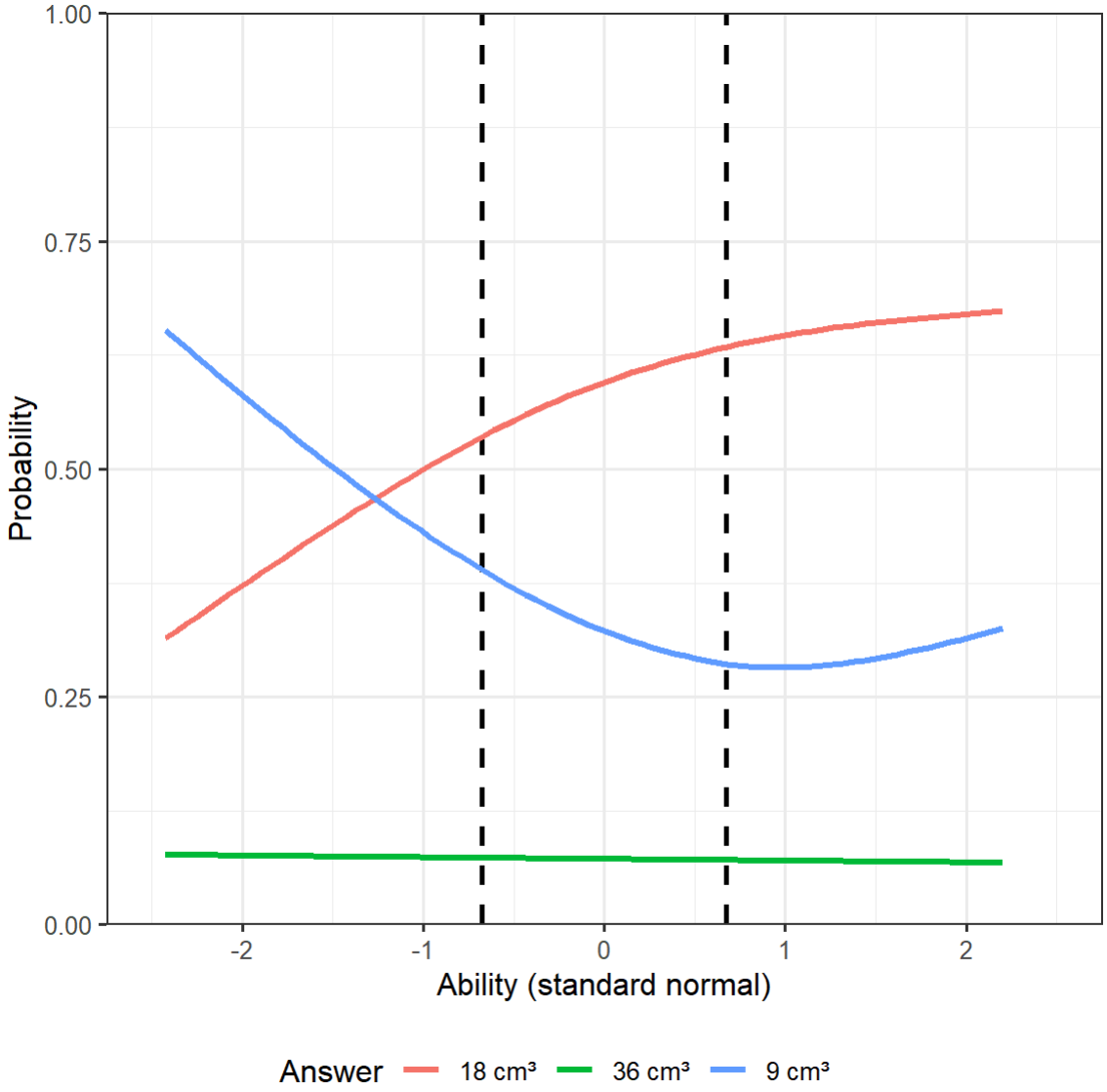
Answer	Frequency	Percent
18 cm^3	146	52.1%
27 cm^3 (correct)	26	9.3%
36 cm^3	19	6.8%
9 cm^3	89	31.8%
Overall	280	100.0%

10.1.13.4 Distractors functioning:

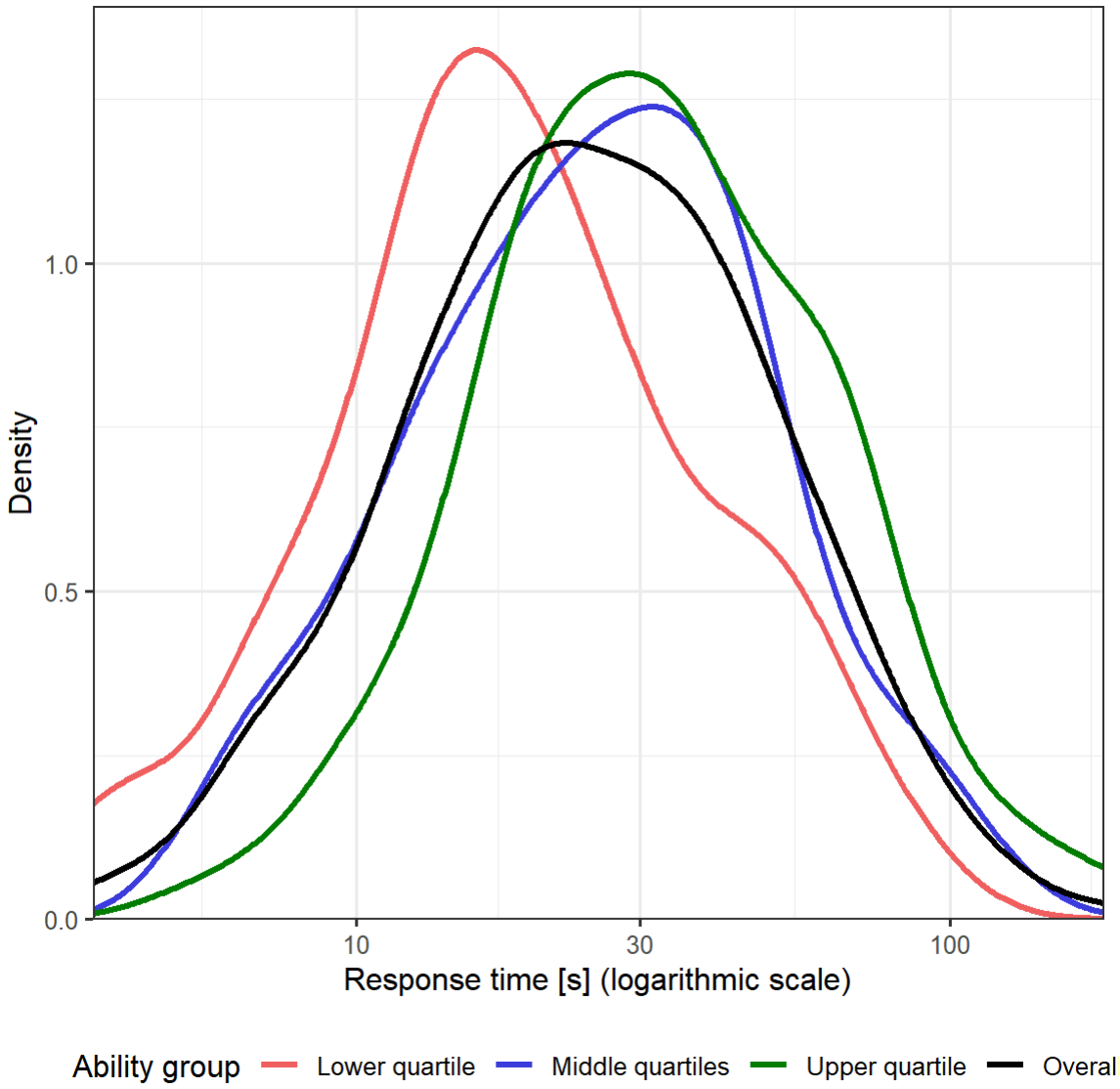
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
18 cm^3	29	72	45	146	45.3%	60.0%	64.3%	57.5%
36 cm^3	5	9	5	19	7.8%	7.5%	7.1%	7.5%
9 cm^3	30	39	20	89	46.9%	32.5%	28.6%	35.0%
Overall	64	120	70	254	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



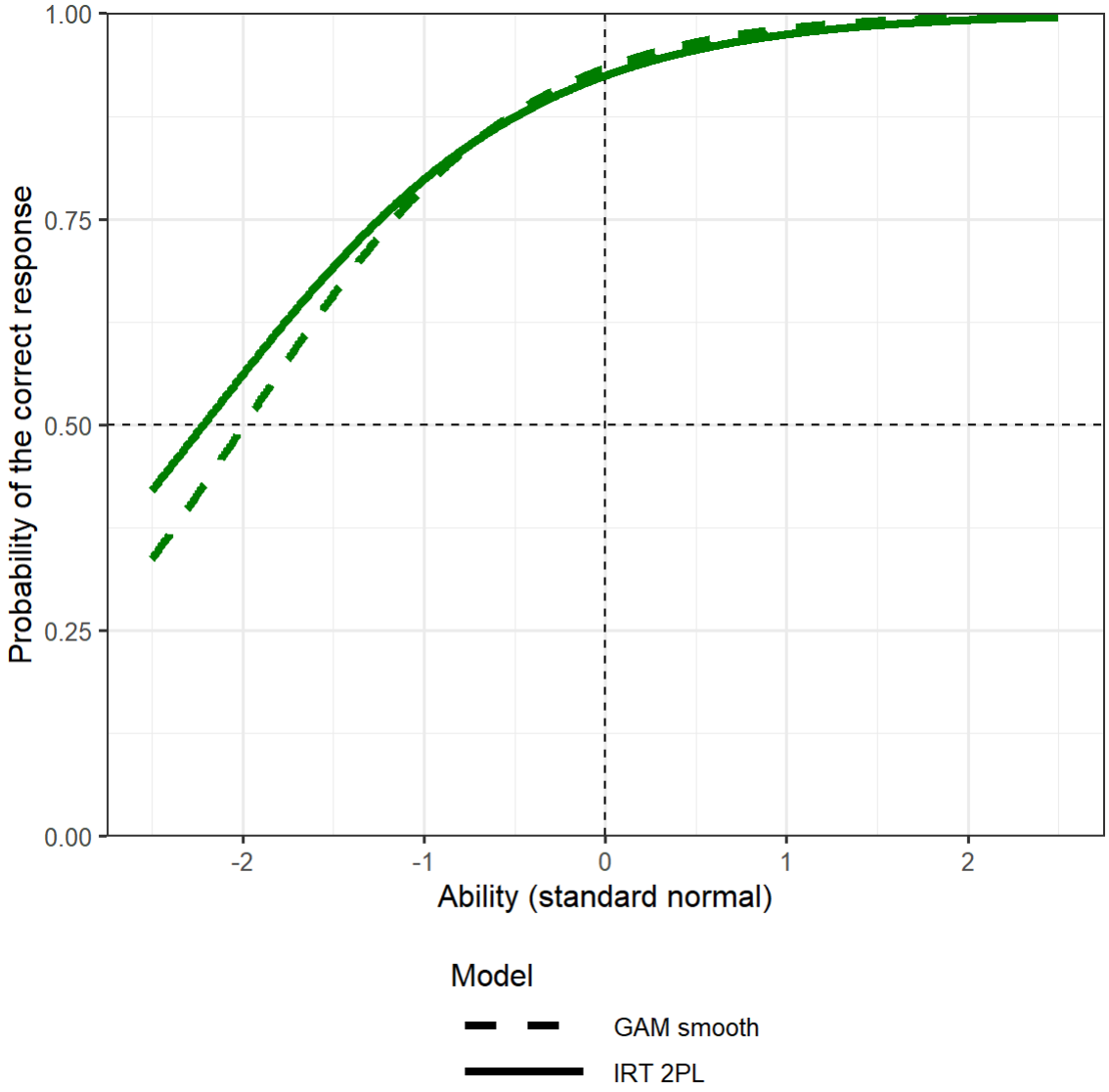
10.1.13.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.6	12.2	17.9	28.2	90.9	2.9	0.7
Middle quartiles	5.6	14.9	25.9	41.1	122.0	3.2	0.7
Upper quartile	5.7	20.5	32.8	52.5	181.4	3.5	0.7
Overall	3.6	14.9	24.6	40.4	181.4	3.2	0.7

10.1.14 M890029 (LLM)

10.1.14.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	88.71%	319	LLM complex setup	MC	0.35	0.25	1.13	−2.22	0.01

10.1.14.2 Gender DIF:

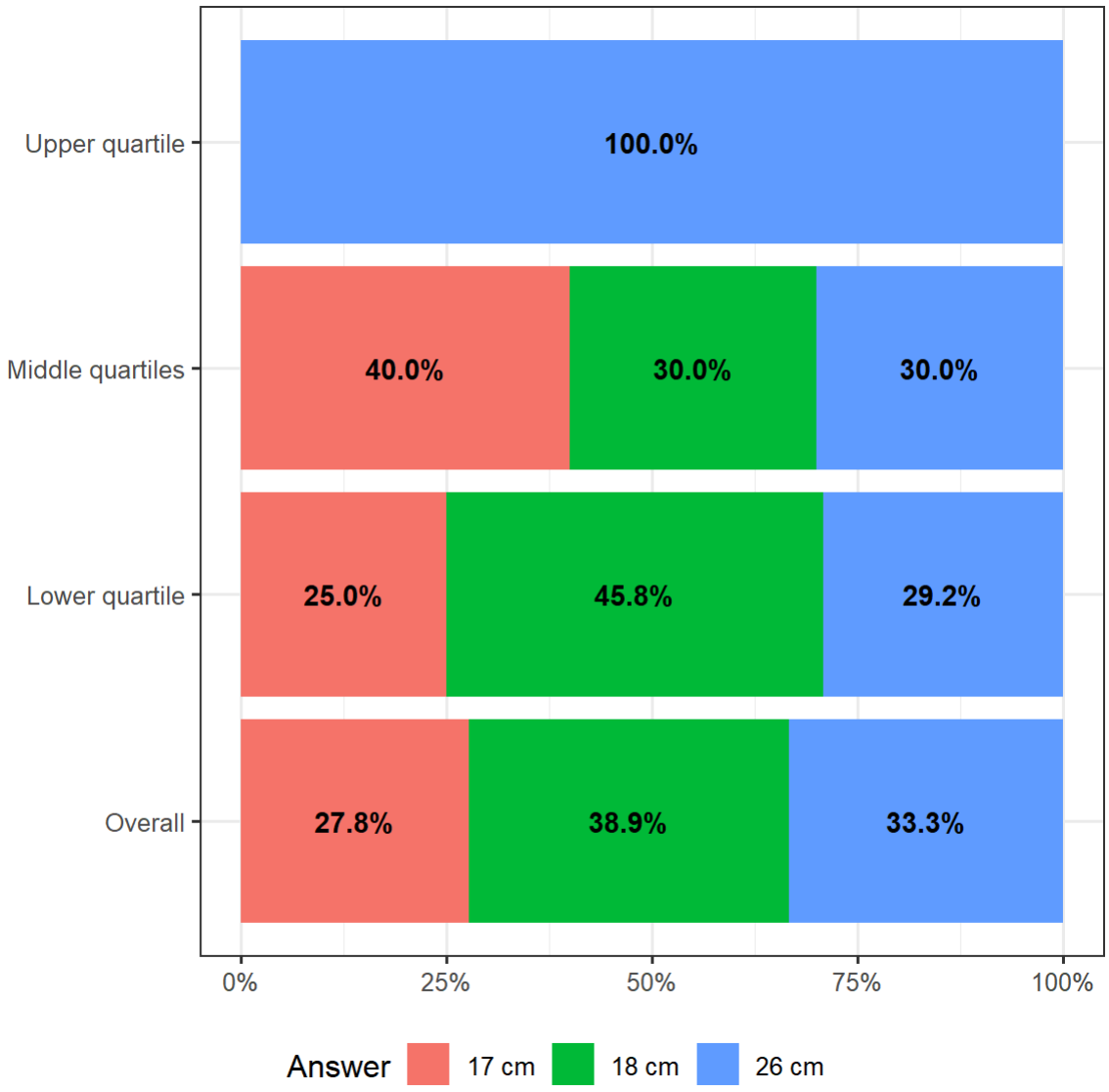
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.77	0.250	1.000	0.013	N/A

10.1.14.3 Frequency of answers:

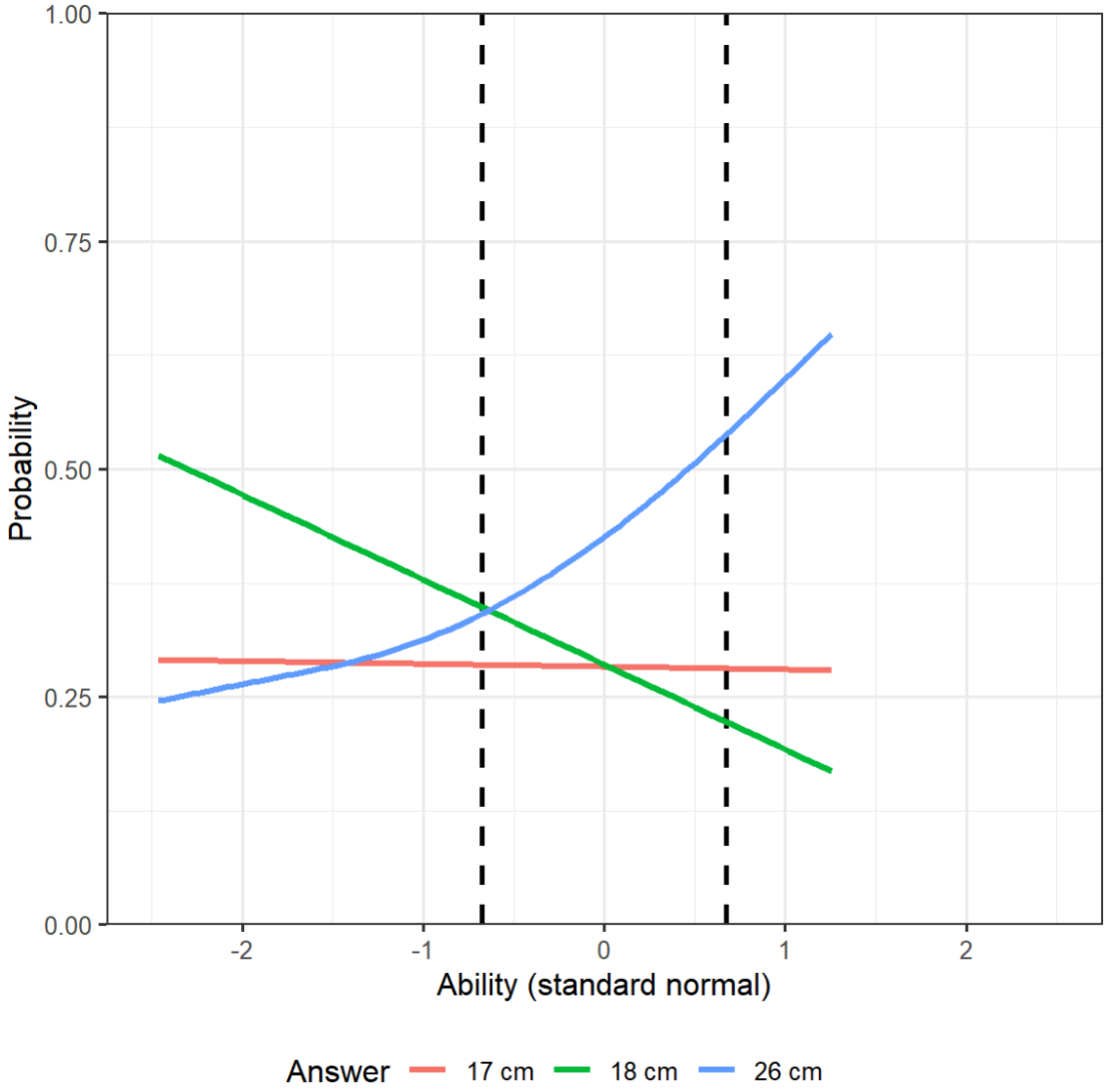
Answer	Frequency	Percent
17 cm	10	3.1%
18 cm	14	4.4%
22 cm (correct)	283	88.7%
26 cm	12	3.8%
Overall	319	100.0%

10.1.14.4 Distractors functioning:

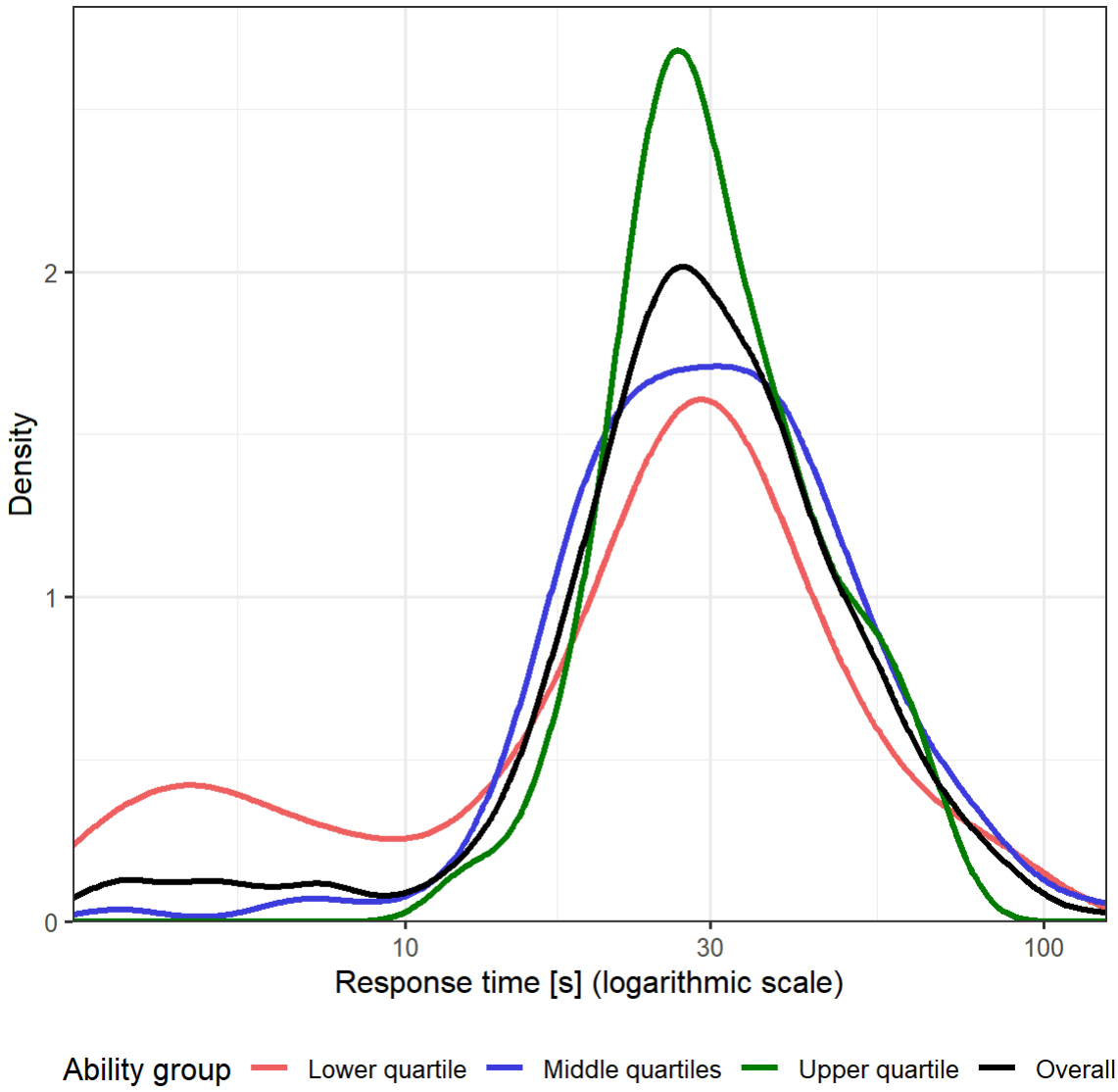
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
17 cm	6	4	0	10	25.0%	40.0%	0.0%	27.8%
18 cm	11	3	0	14	45.8%	30.0%	0.0%	38.9%
26 cm	7	3	2	12	29.2%	30.0%	100.0%	33.3%
Overall	24	10	2	36	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



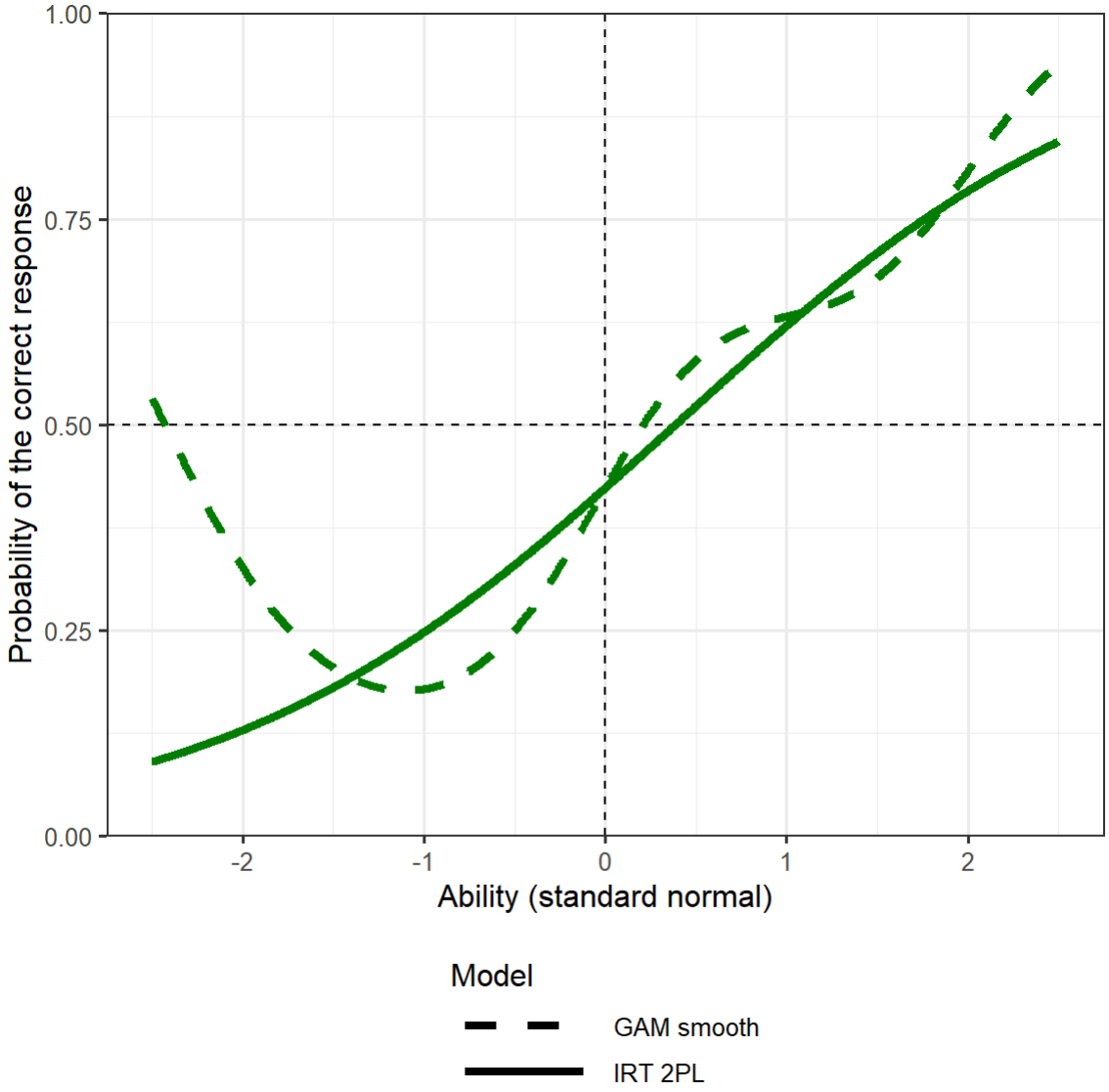
10.1.14.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	15.6	26.5	35.1	90.5	3.1	0.8
Middle quartiles	3.6	21.6	30.1	42.0	125.0	3.4	0.5
Upper quartile	12.4	24.3	28.6	39.2	67.8	3.4	0.4
Overall	3.0	22.0	28.6	39.7	125.0	3.3	0.6

10.1.15 M890026 (LLM)

10.1.15.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	43.57%	319	LLM complex setup	MC	0.36	0.21	0.80	0.39	0.00

10.1.15.2 Gender DIF:

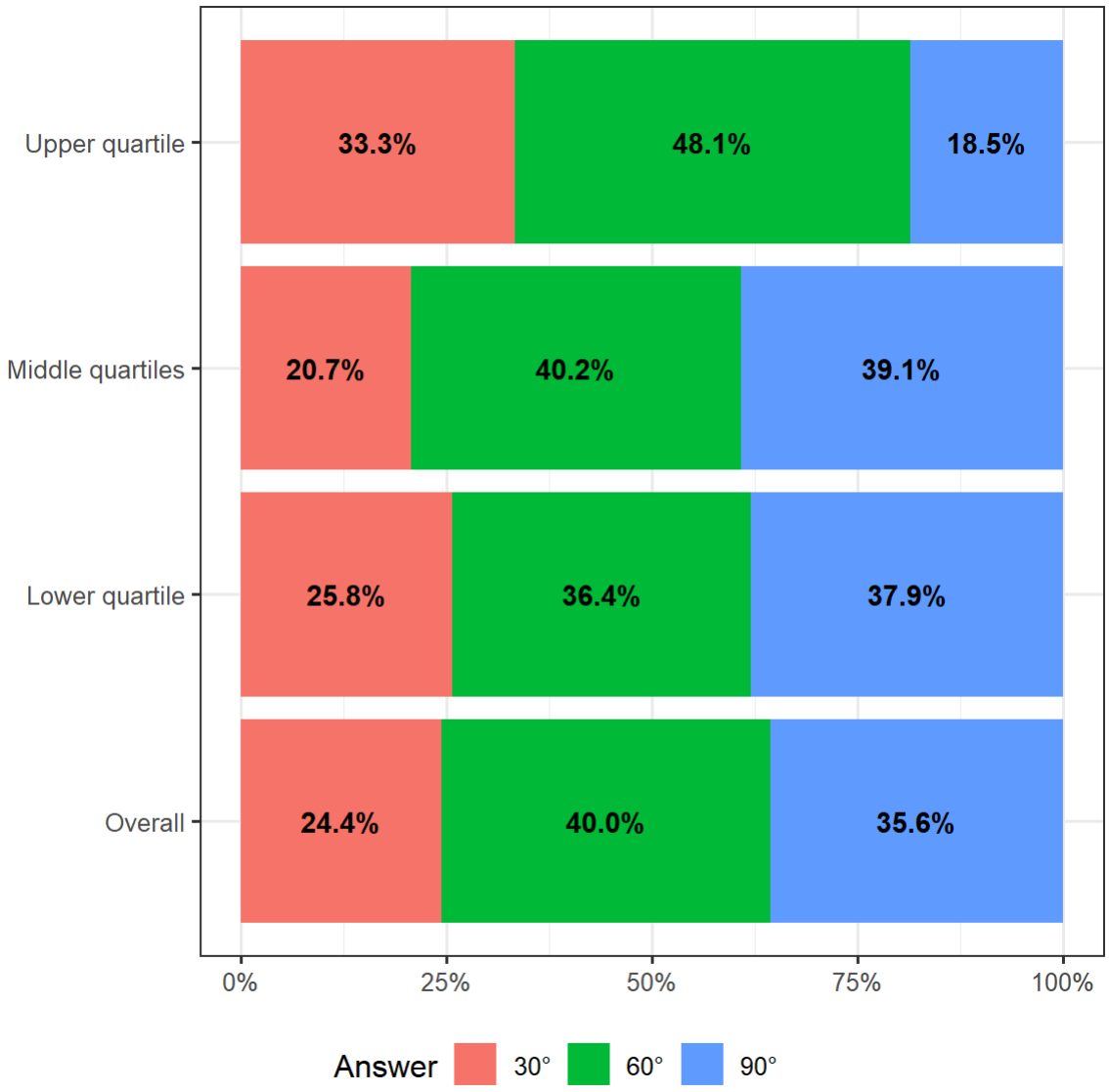
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.86	0.651	1.000	0.002	N/A

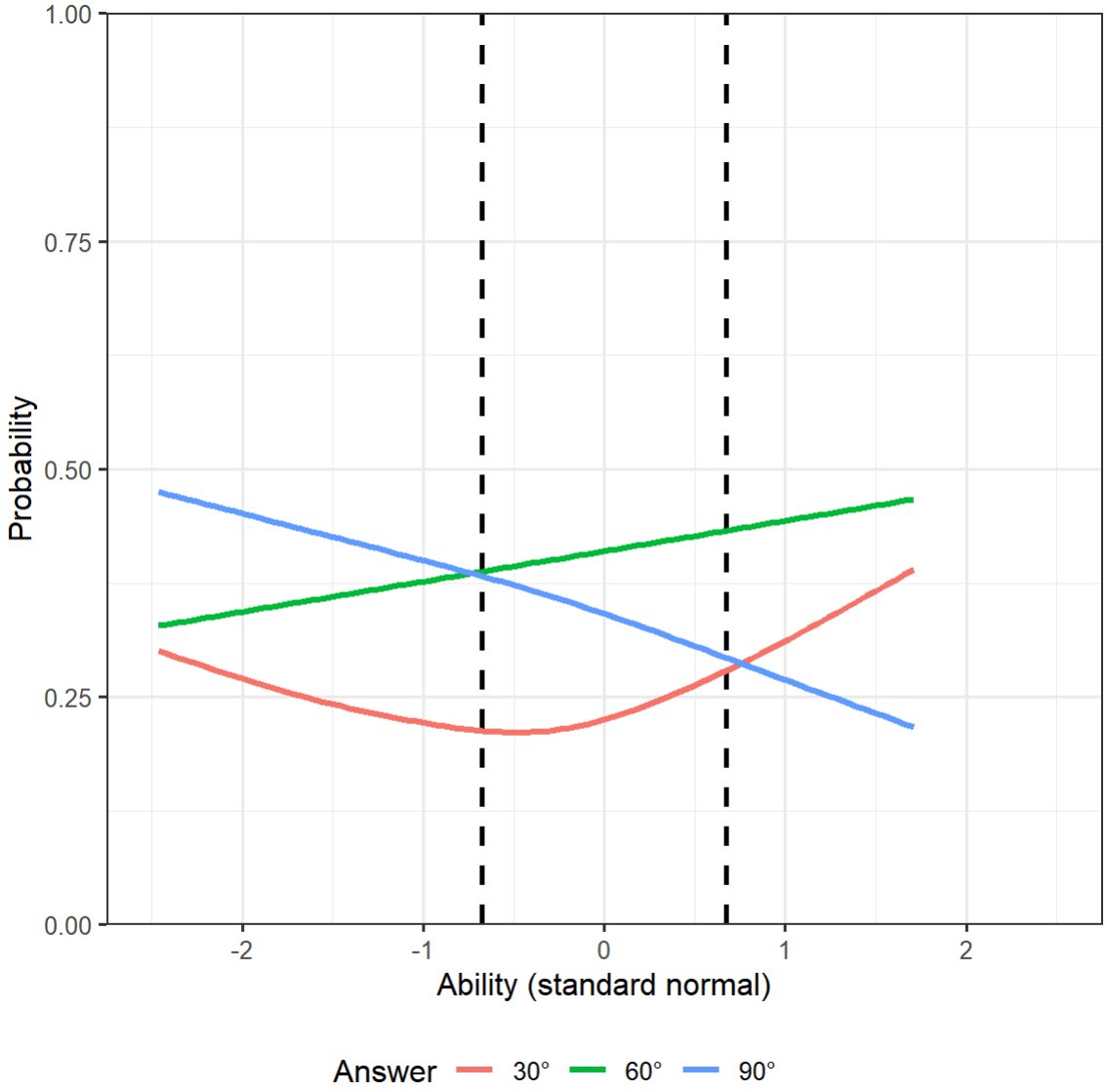
10.1.15.3 Frequency of answers:

Answer	Frequency	Percent
120° (correct)	139	43.6%
30°	44	13.8%
60°	72	22.6%
90°	64	20.1%
Overall	319	100.0%

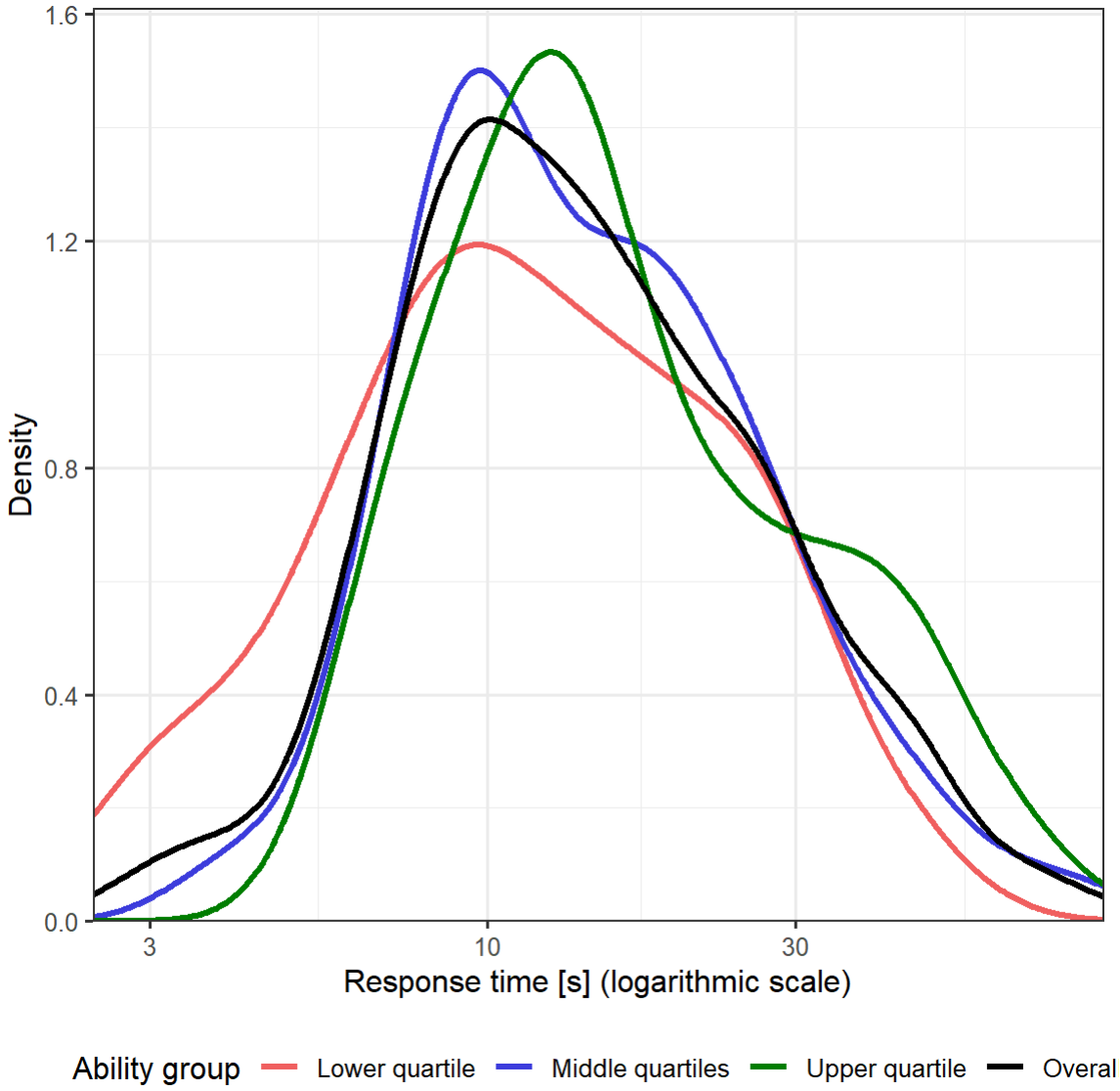
10.1.15.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
30°	17	18	9	44	25.8%	20.7%	33.3%	24.4%
60°	24	35	13	72	36.4%	40.2%	48.1%	40.0%
90°	25	34	5	64	37.9%	39.1%	18.5%	35.6%
Overall	66	87	27	180	100.0%	100.0%	100.0%	100.0%


Upper quartile
Middle quartiles
Lower quartile
Overall
33.3%
48.1%
18.5%
20.7%
40.2%
39.1%
25.8%
36.4%
37.9%
24.4%
40.0%
35.6%
0%
25%
50%
75%
100%
Answer
30°
60°
90°



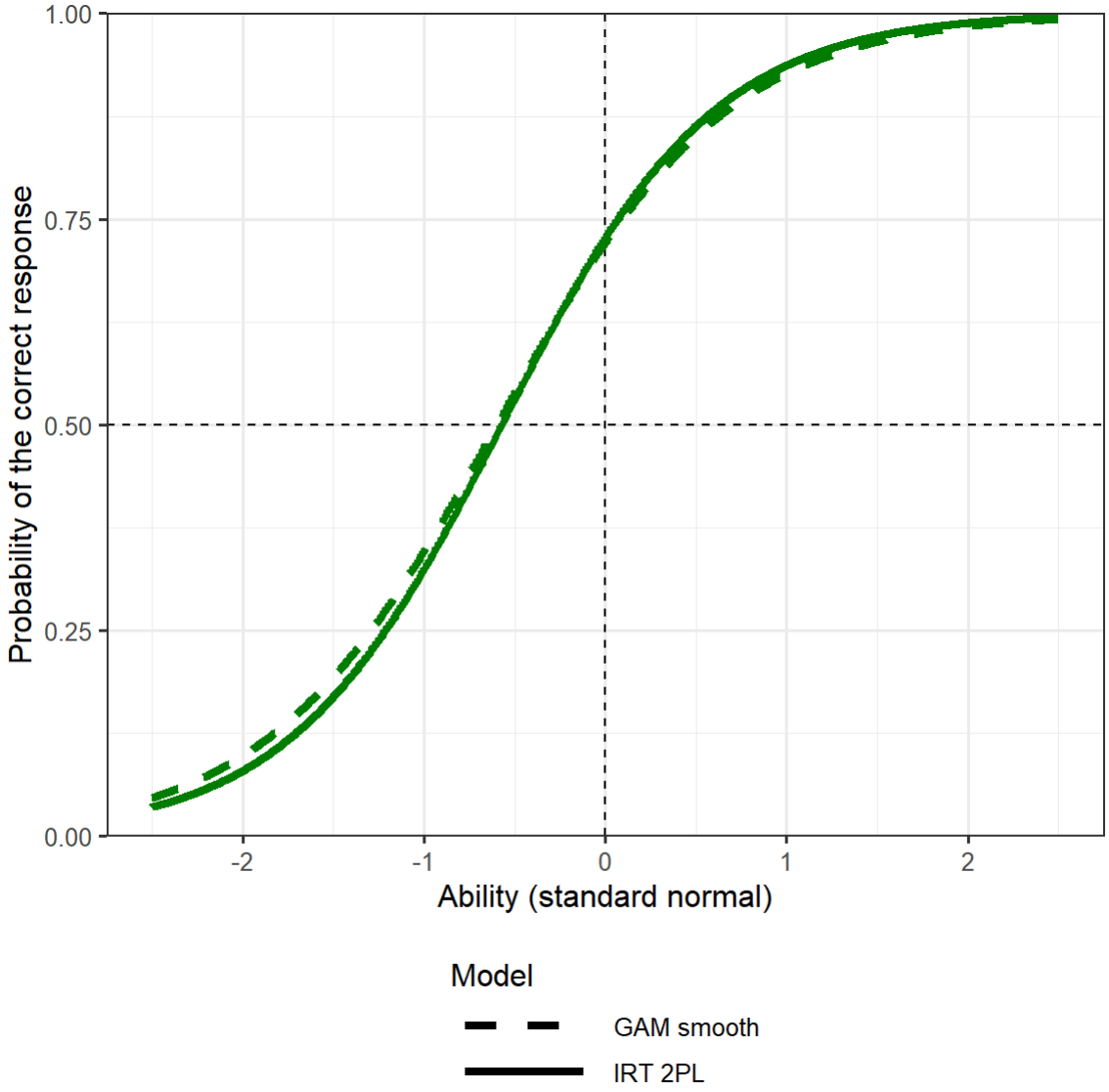
10.1.15.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.4	7.3	11.2	19.5	45.7	2.4	0.7
Middle quartiles	3.6	9.1	13.4	21.5	89.9	2.7	0.6
Upper quartile	6.2	9.9	13.8	25.1	74.5	2.8	0.6
Overall	2.4	8.9	13.2	21.5	89.9	2.6	0.6

10.1.16 M090051 (LLM)

10.1.16.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	65.10%	299	LLM teachers' selection	MC	0.53	0.40	1.72	−0.57	0.0(

10.1.16.2 Gender DIF:

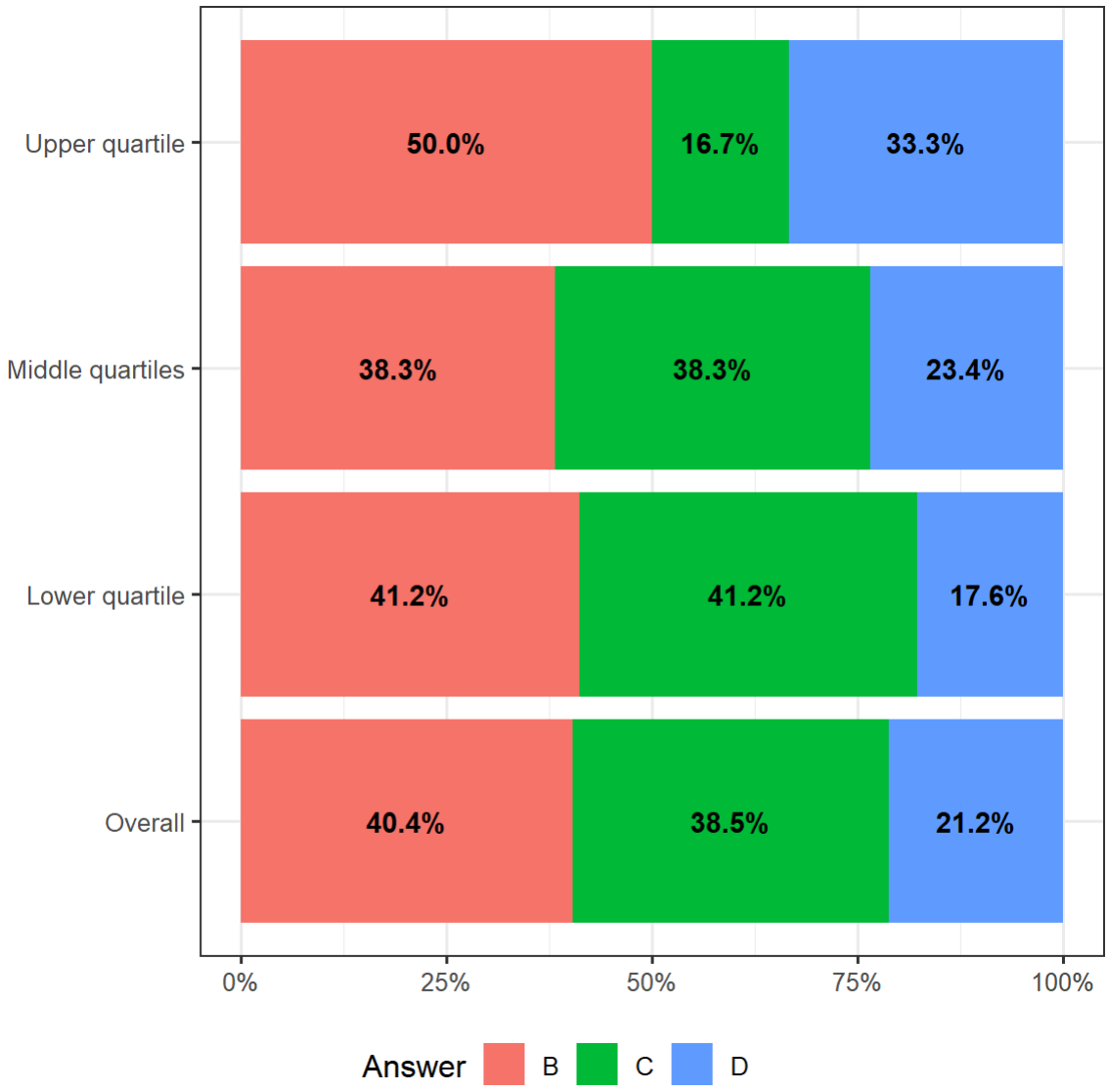
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.48	0.785	1.000	0.001	N/A

10.1.16.3 Frequency of answers:

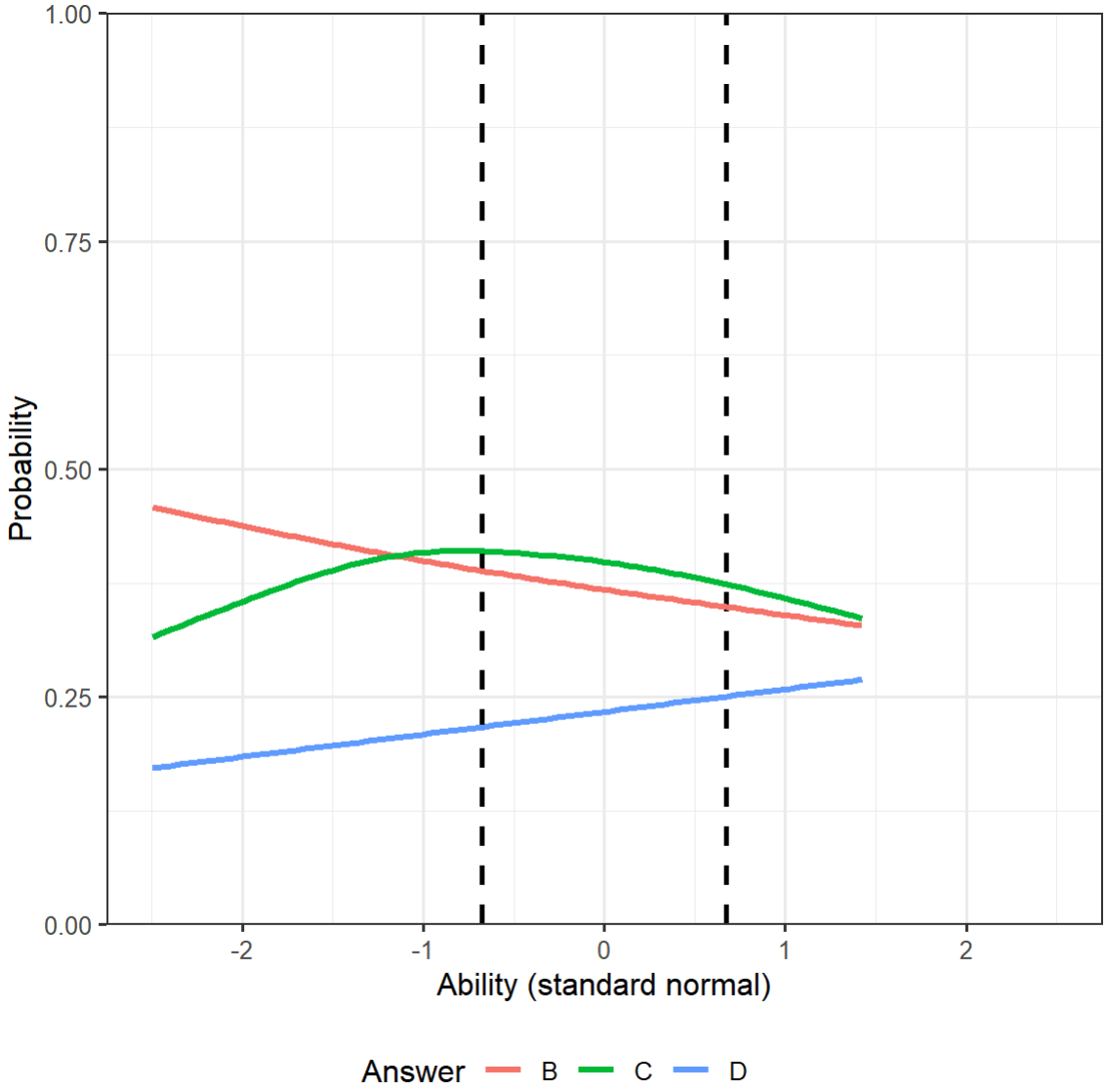
Answer	Frequency	Percent
A (correct)	194	65.1%
B	42	14.1%
C	40	13.4%
D	22	7.4%
Overall	298	100.0%

10.1.16.4 Distractors functioning:

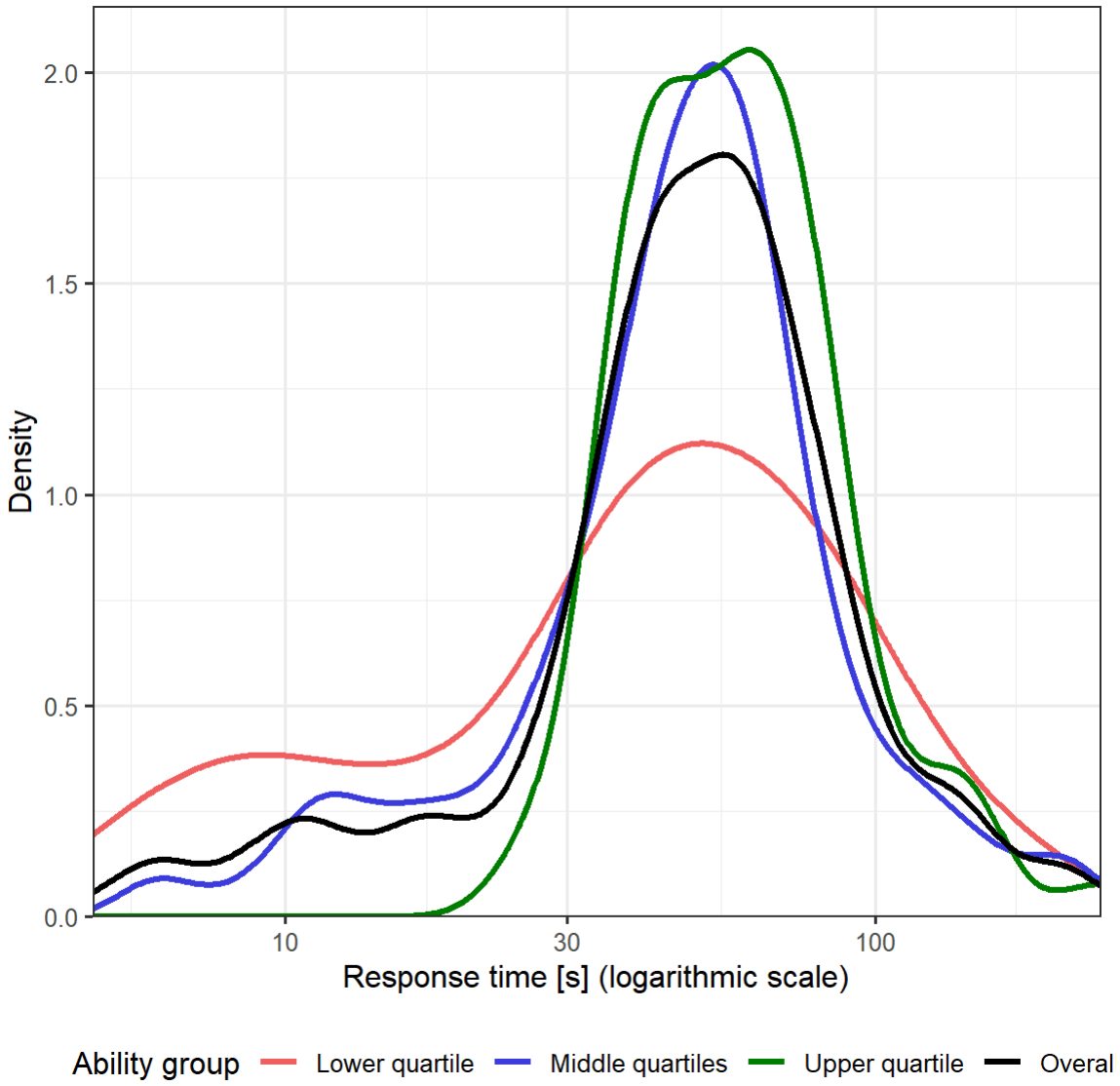
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
B	21	18	3	42	41.2%	38.3%	50.0%	40.4%
C	21	18	1	40	41.2%	38.3%	16.7%	38.5%
D	9	11	2	22	17.6%	23.4%	33.3%	21.2%
Overall	51	47	6	104	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



10.1.16.5 Answering time:

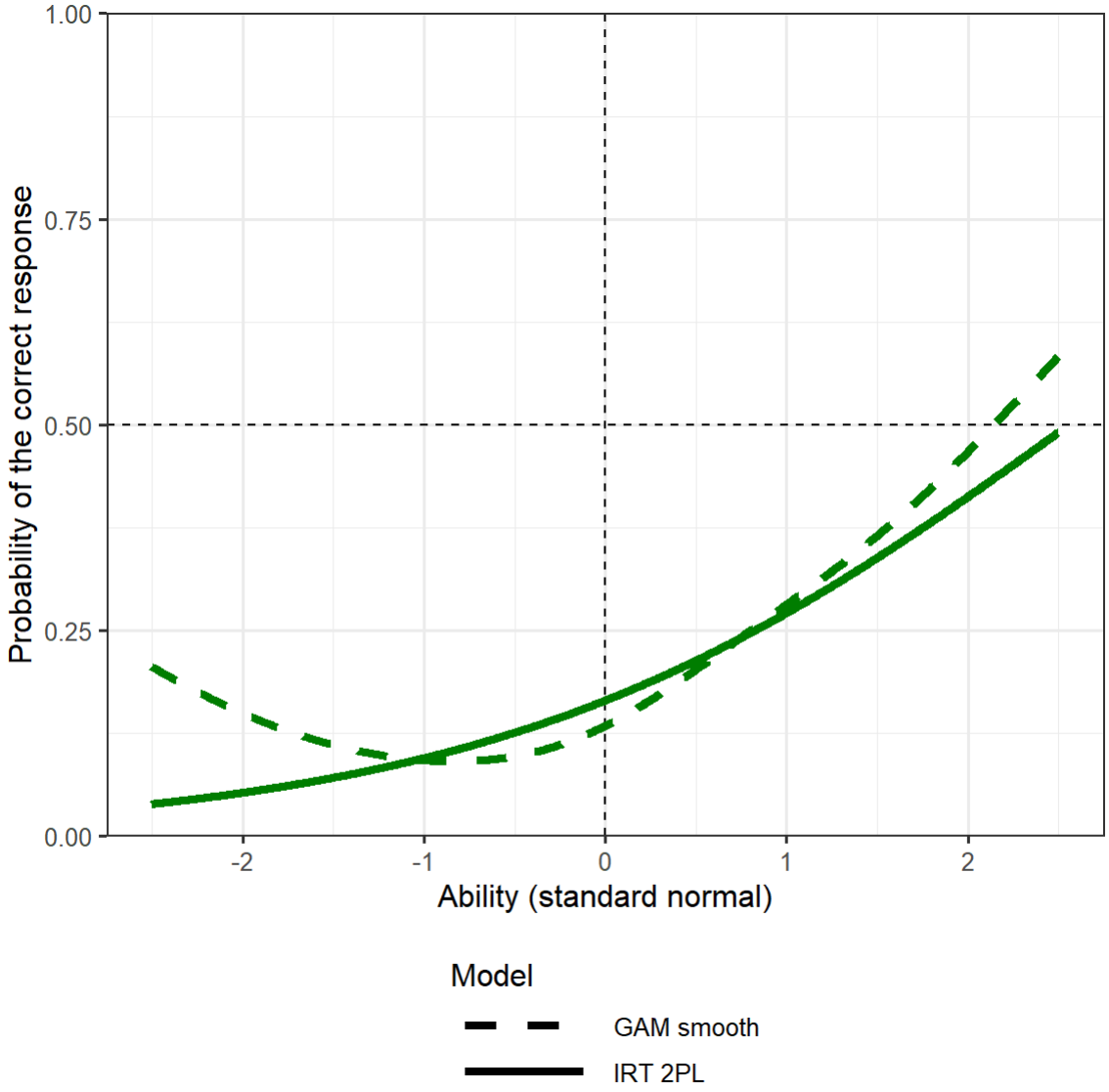


	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.7	22.1	42.9	70.8	190.7	3.6	0.9
Middle quartiles	6.0	35.4	48.6	64.4	215.1	3.8	0.7
Upper quartile	25.6	41.8	56.3	73.1	240.5	4.1	0.4
Overall	4.7	35.6	51.0	68.4	240.5	3.8	0.7

10.1.17 M090044 (LLM)

10.1.17.1 Item fit:

```
## Warning in newton(lsp = lsp, X = G$X, y = G$y, Eb = G$Eb, UrS = G$UrS, L = G$L,
## : Fitting terminated with step failure - check results carefully
## Warning in newton(lsp = lsp, X = G$X, y = G$y, Eb = G$Eb, UrS = G$UrS, L = G$L,
## : Fitting terminated with step failure - check results carefully
```



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	18.12%	299	LLM teachers' selection	MC	0.22	0.18	0.63	2.56	0.03

10.1.17.2 Gender DIF:

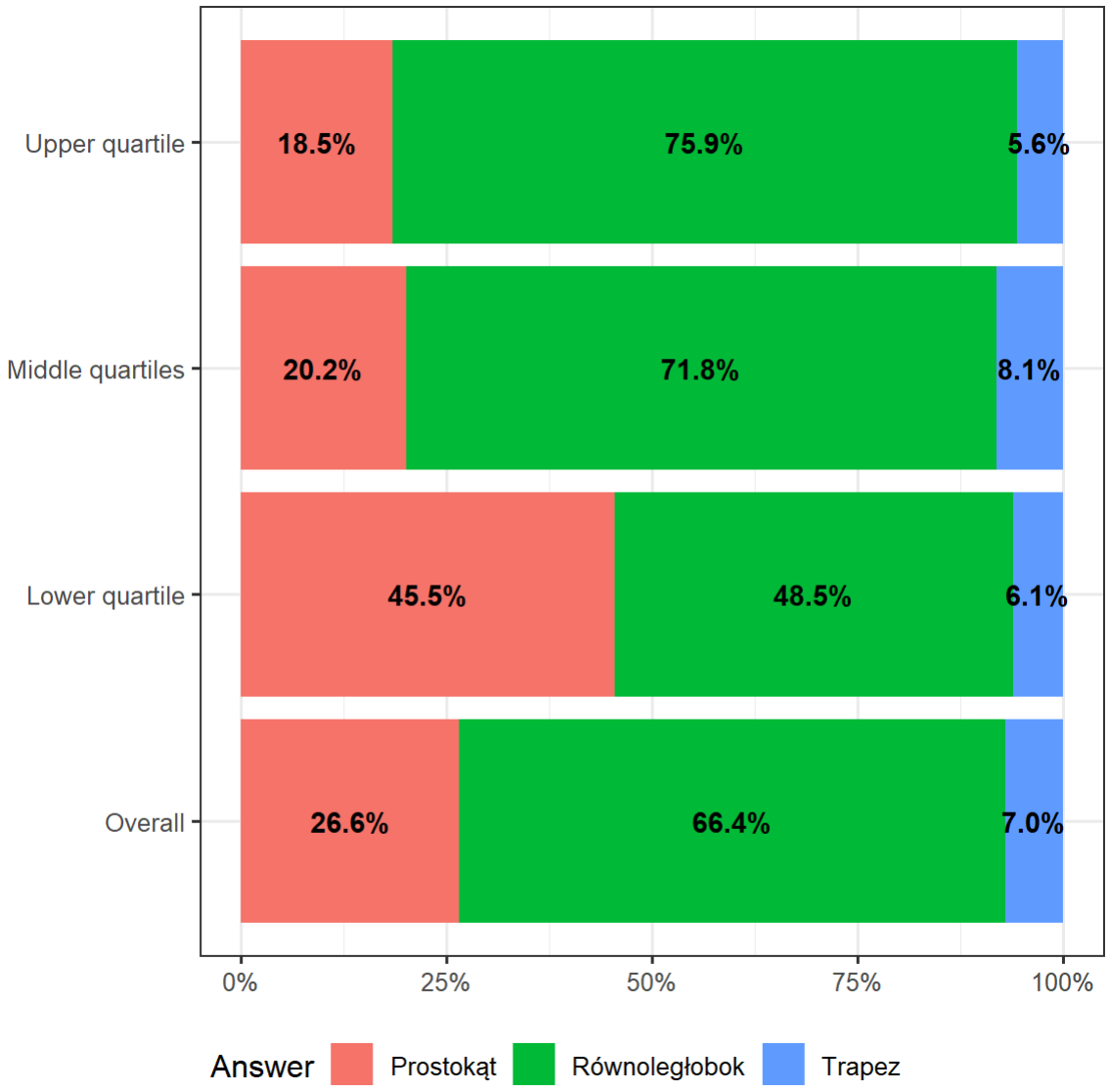
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.69	0.708	1.000	0.003	N/A

10.1.17.3 Frequency of answers:

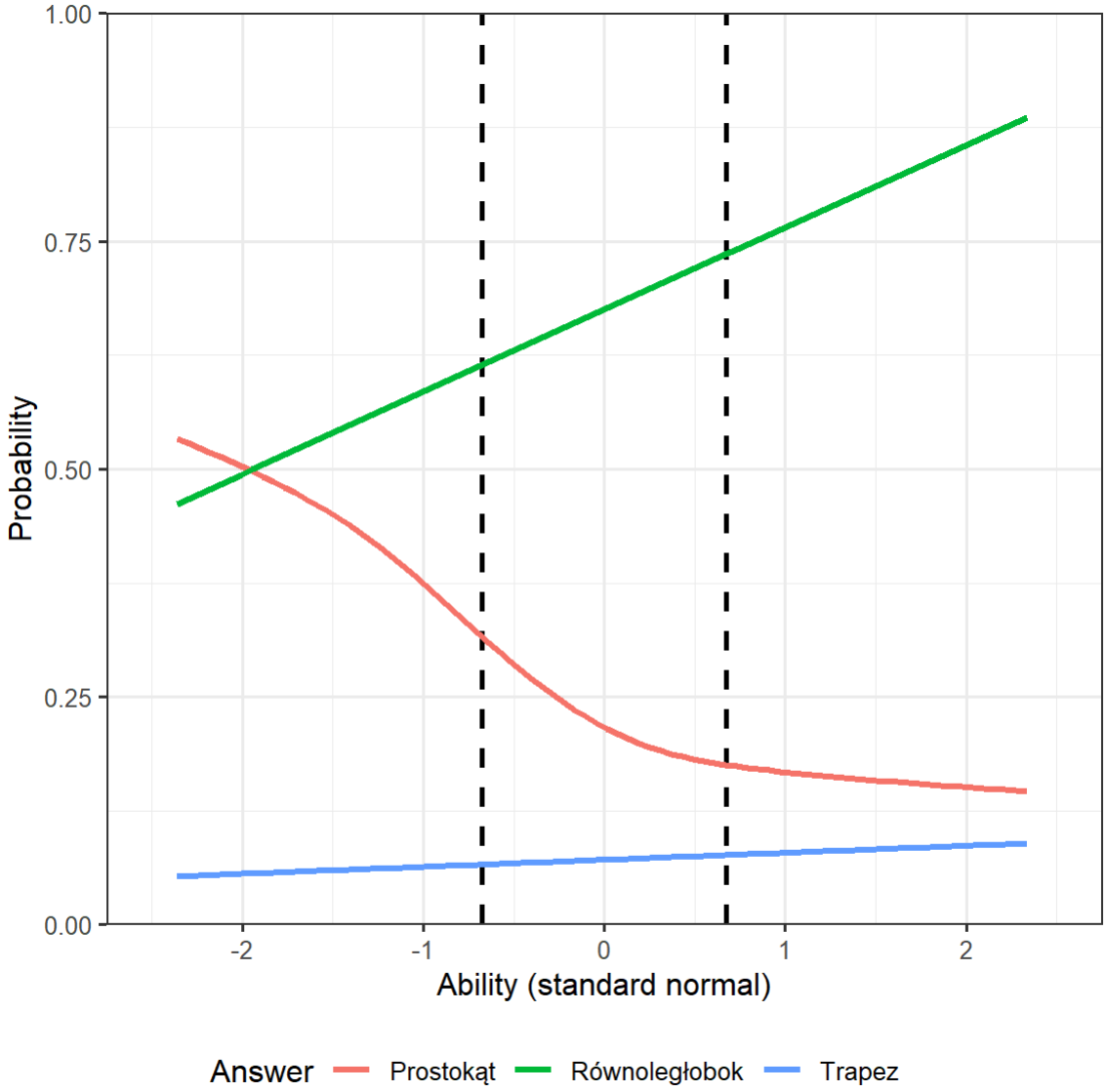
Answer	Frequency	Percent
Prostokąt	65	21.8%
Romb (correct)	54	18.1%
Równoległobok	162	54.4%
Trapez	17	5.7%
Overall	298	100.0%

10.1.17.4 Distractors functioning:

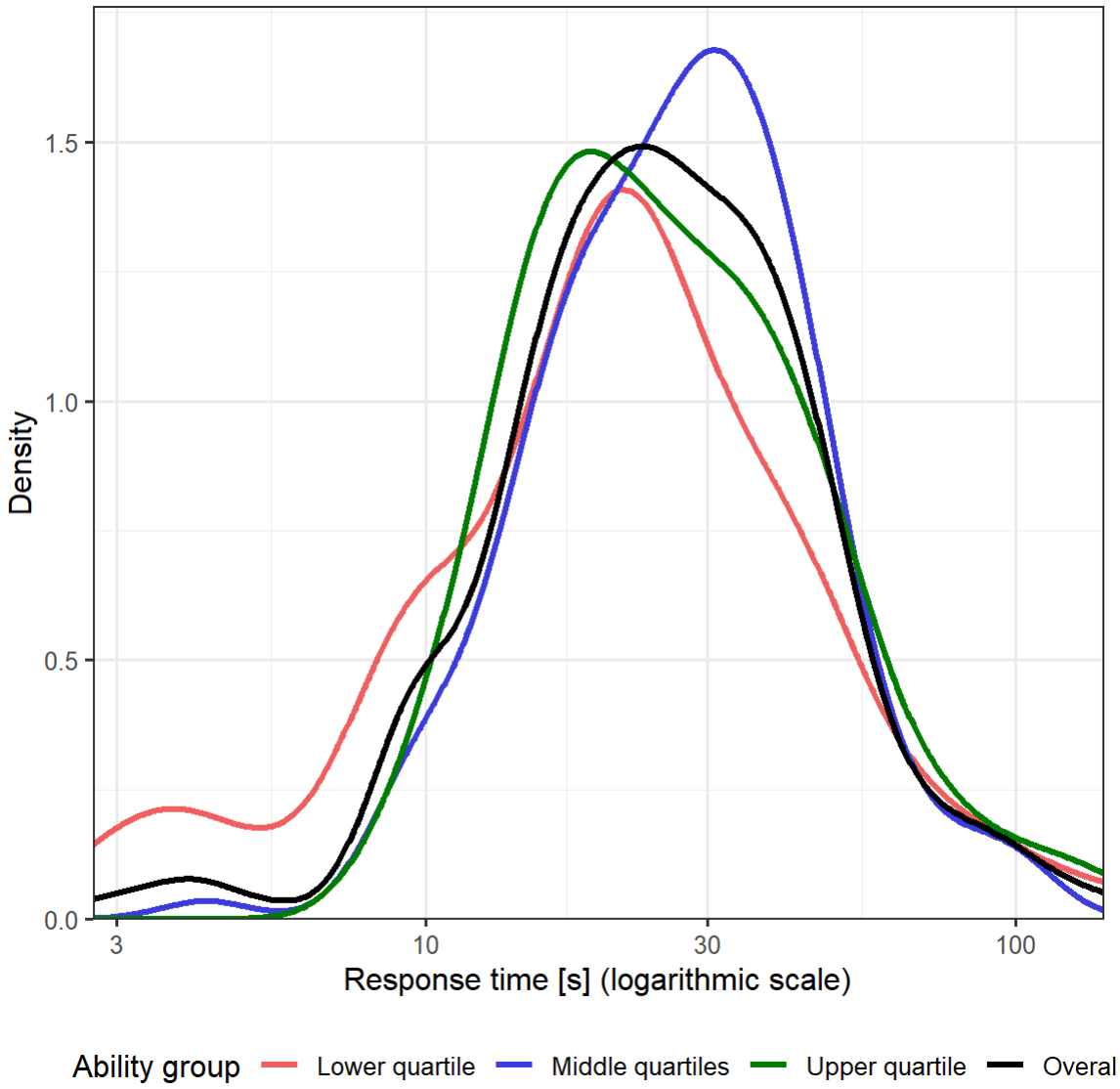
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Prostokąt	30	25	10	65	45.5%	20.2%	18.5%	26.6%
Równoległobok	32	89	41	162	48.5%	71.8%	75.9%	66.4%
Trapez	4	10	3	17	6.1%	8.1%	5.6%	7.0%
Overall	66	124	54	244	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



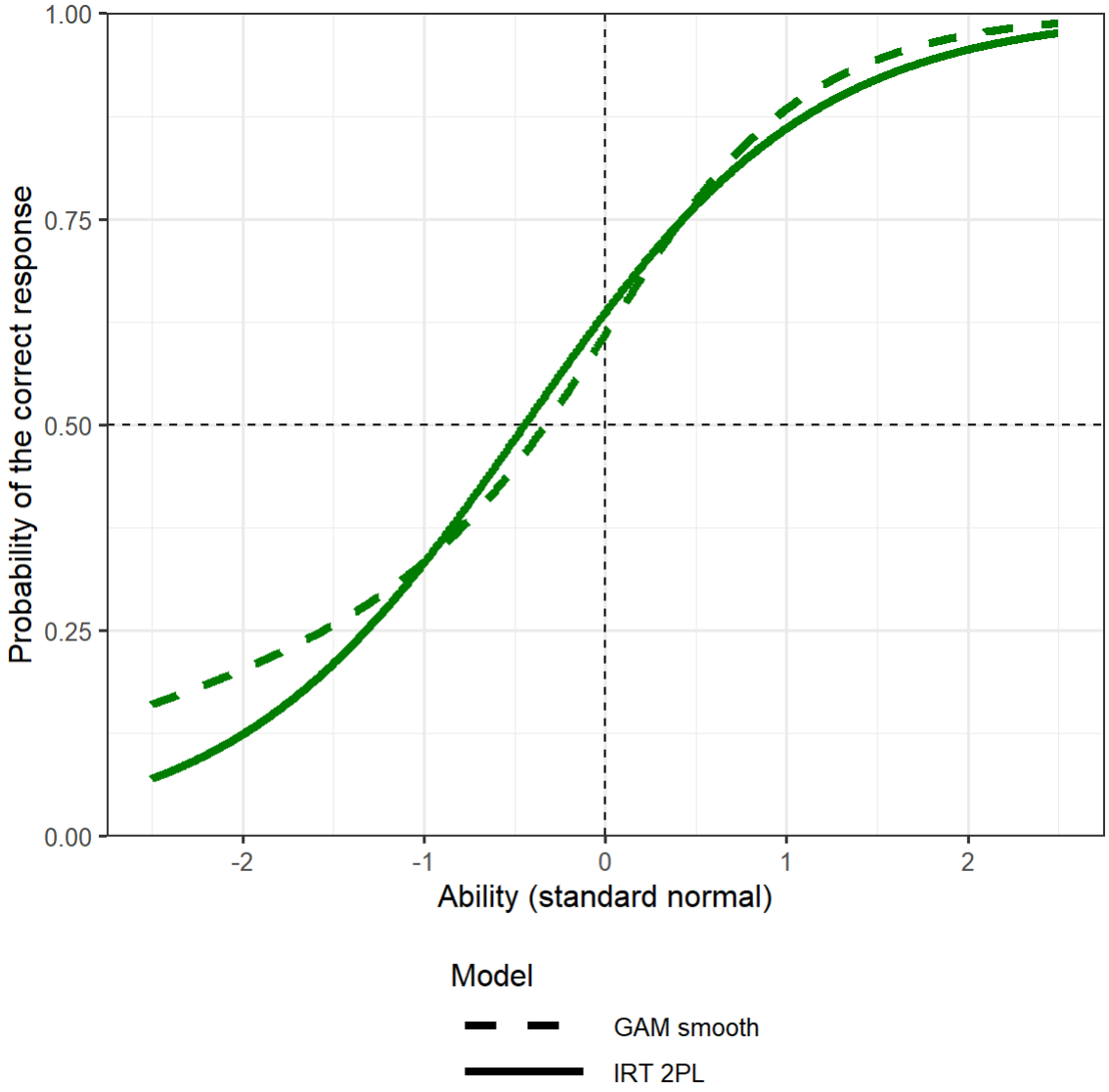
10.1.17.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	13.9	21.4	33.8	140.5	3.0	0.8
Middle quartiles	4.2	18.4	26.9	38.3	105.5	3.3	0.5
Upper quartile	9.5	16.7	24.6	36.8	135.7	3.3	0.6
Overall	2.7	16.8	24.8	36.9	140.5	3.2	0.6

10.1.18 M090016 (LLM)

10.1.18.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS generated items: Discrimination	Item difficulty	RMSE
1	60.36%	281	LLM teachers' selection	MC	0.48	0.29	1.26	−0.44	0.0(

10.1.18.2 Gender DIF:

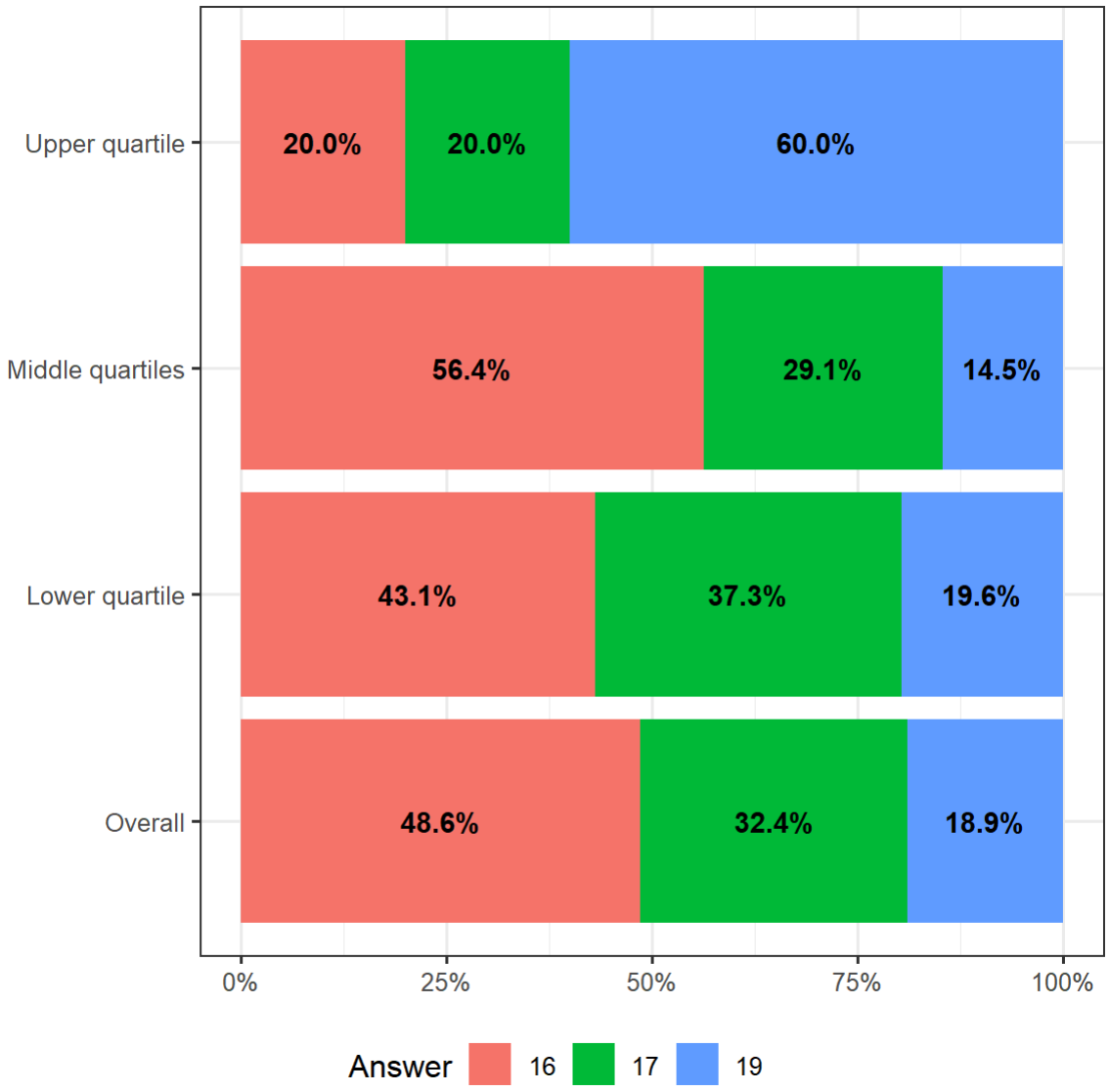
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
8.18	0.017	1.000	0.024	N/A

10.1.18.3 Frequency of answers:

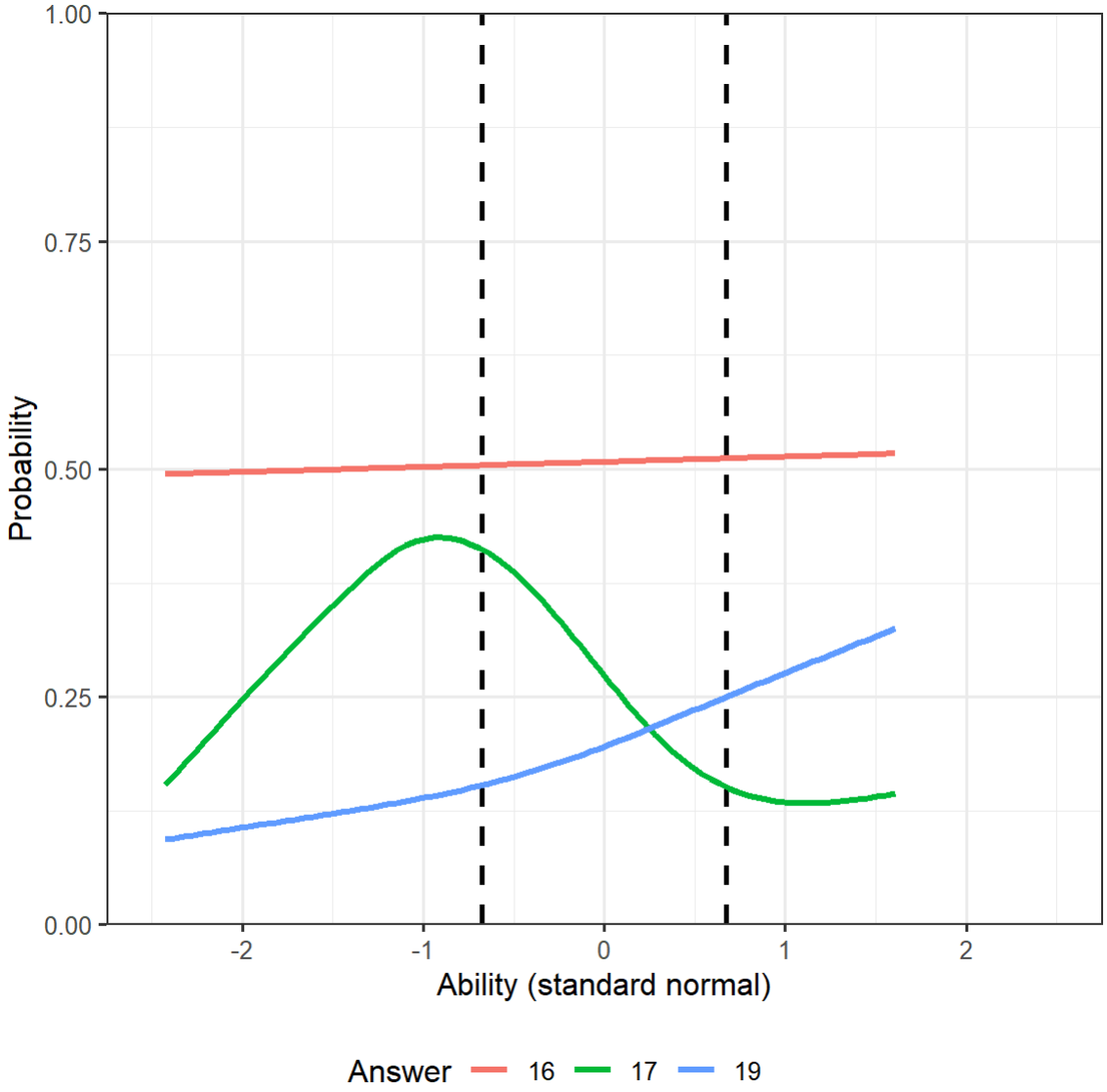
Answer	Frequency	Percent
16	54	19.3%
17	36	12.9%
18 (correct)	169	60.4%
19	21	7.5%
Overall	280	100.0%

10.1.18.4 Distractors functioning:

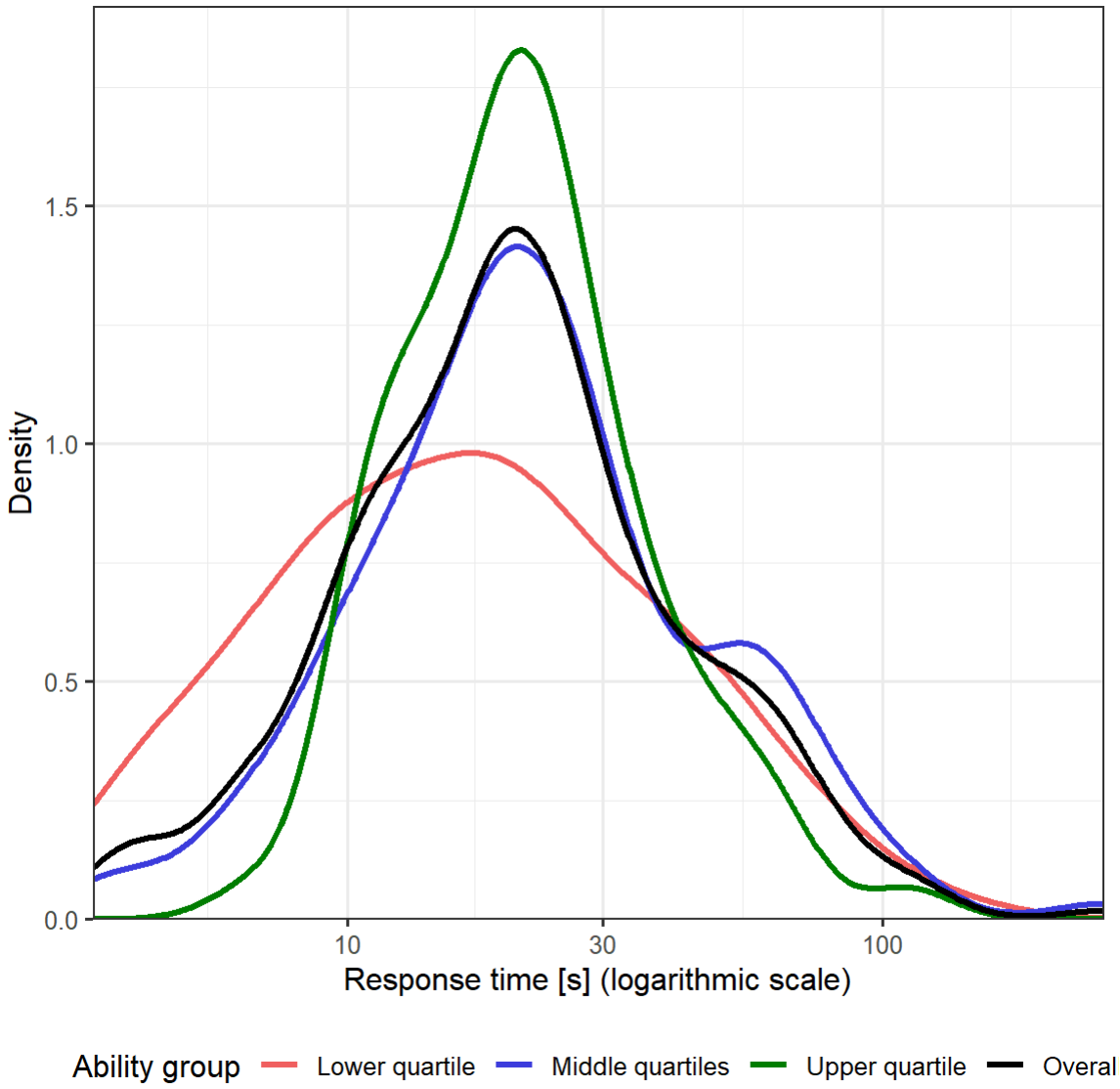
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
16	22	31	1	54	43.1%	56.4%	20.0%	48.6%
17	19	16	1	36	37.3%	29.1%	20.0%	32.4%
19	10	8	3	21	19.6%	14.5%	60.0%	18.9%
Overall	51	55	5	111	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



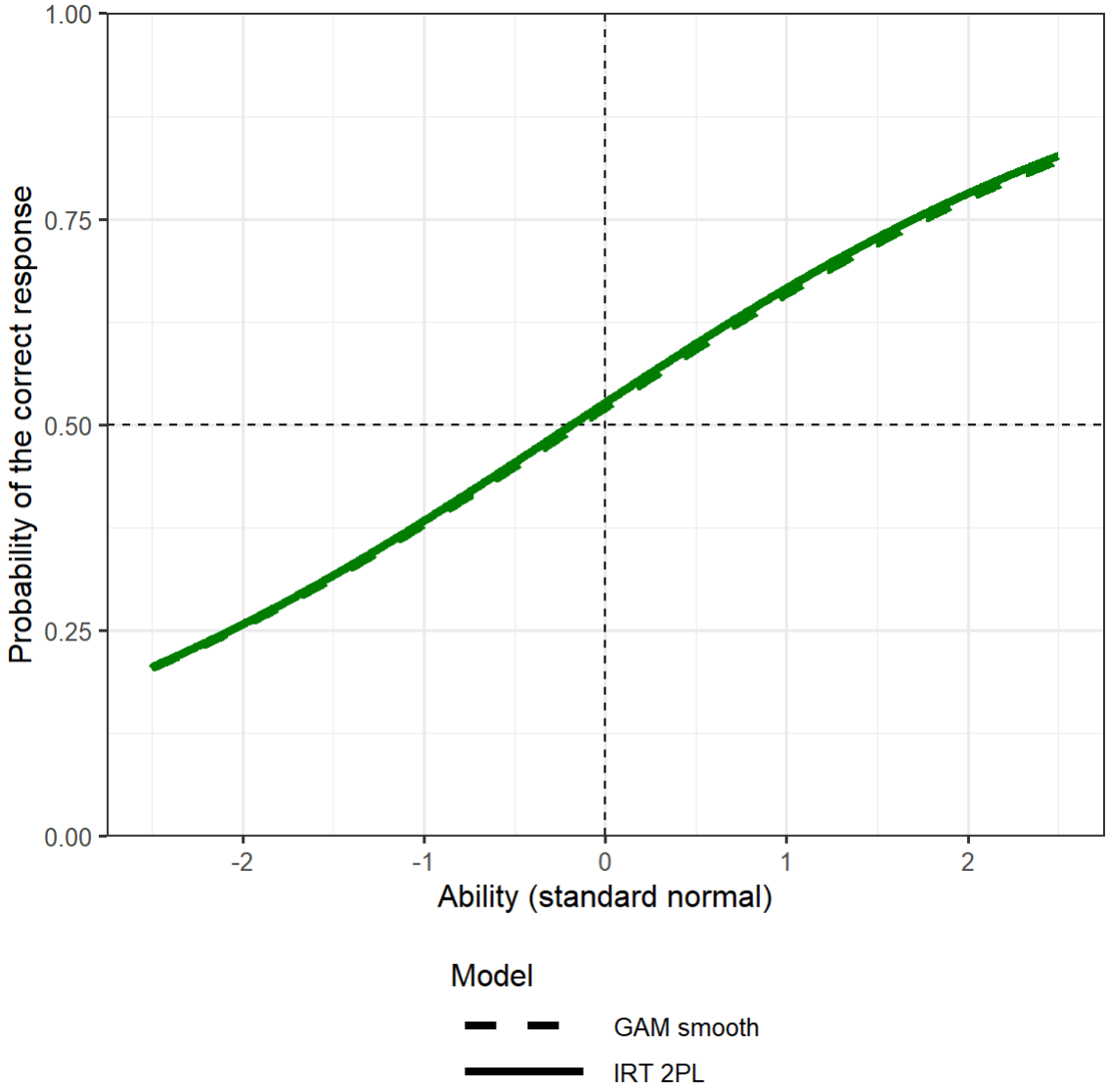
10.1.18.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.6	9.4	17.3	29.2	124.6	2.8	0.8
Middle quartiles	3.4	13.9	20.9	32.8	258.3	3.1	0.7
Upper quartile	6.6	14.9	21.3	29.1	111.8	3.1	0.5
Overall	3.4	12.6	20.1	29.9	258.3	3.0	0.7

10.1.19 M041333 (TIMSS)

10.1.19.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	51.97%	304	TIMSS	MC	0.27	0.27	0.58	−0.18	0.009

10.1.19.2 Gender DIF:

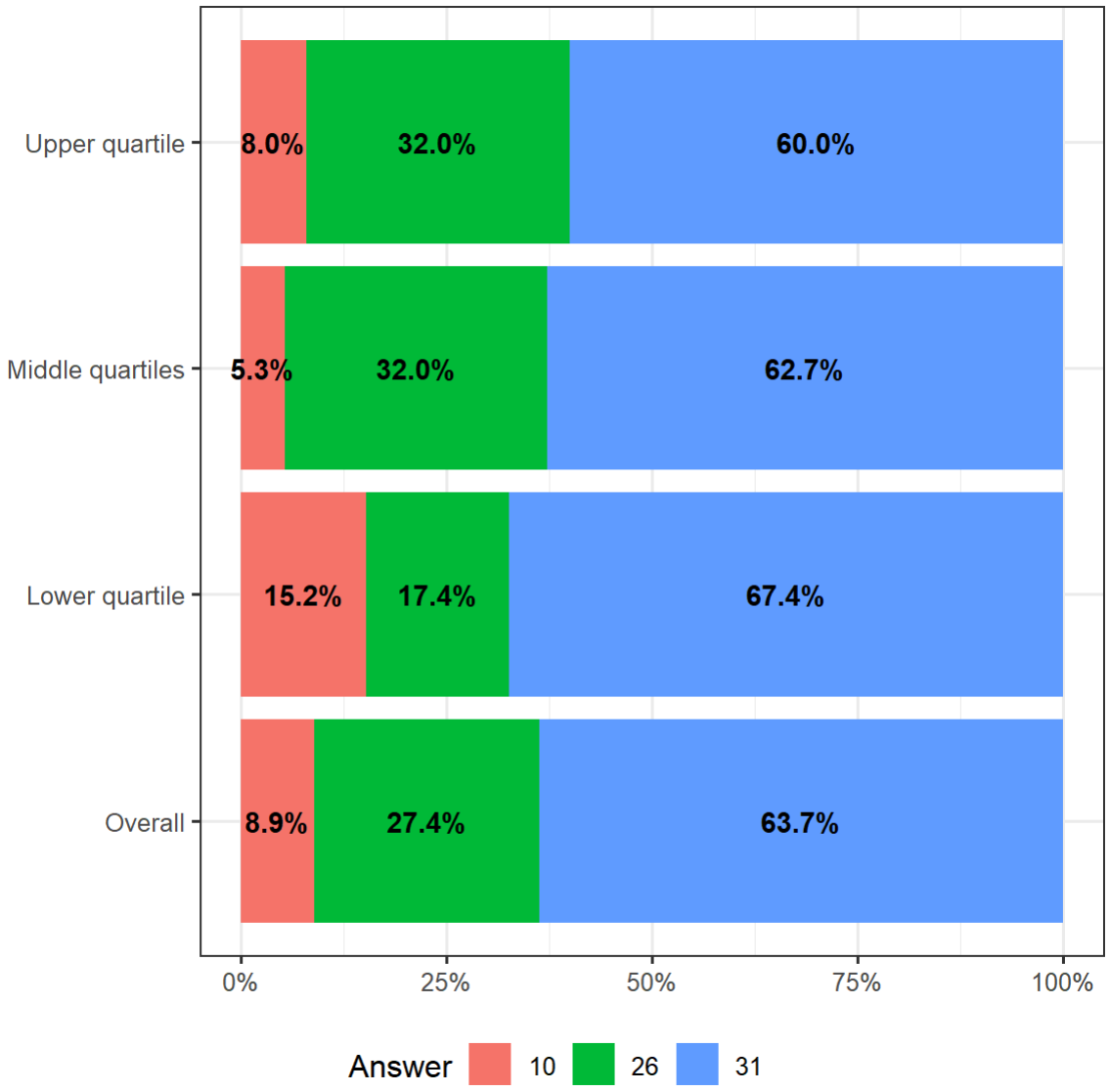
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
8.70	0.013	1.000	0.024	N/A

10.1.19.3 Frequency of answers:

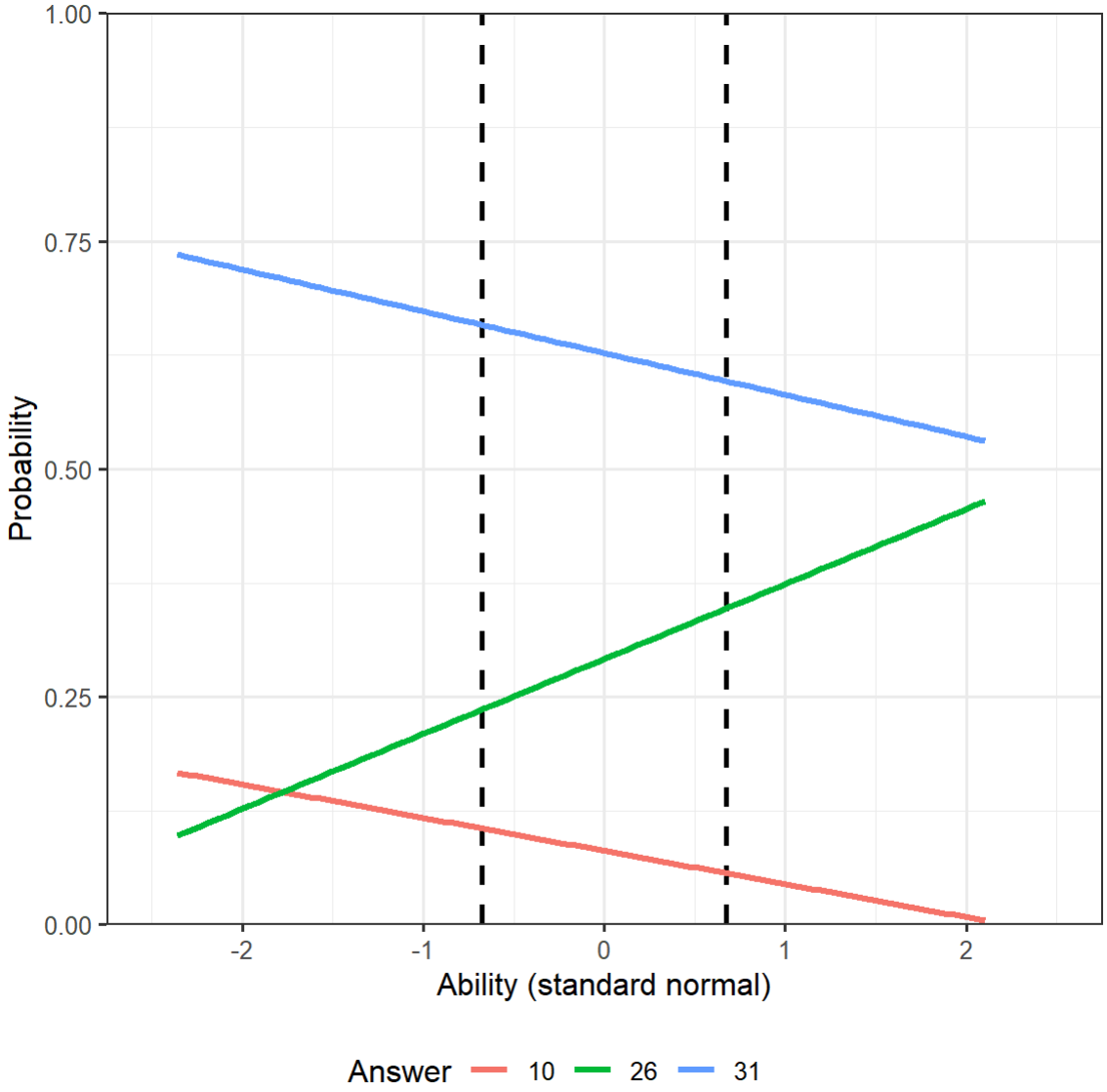
Answer	Frequency	Percent
10	13	4.3%
26	40	13.2%
30 (correct)	158	52.0%
31	93	30.6%
Overall	304	100.0%

10.1.19.4 Distractors functioning:

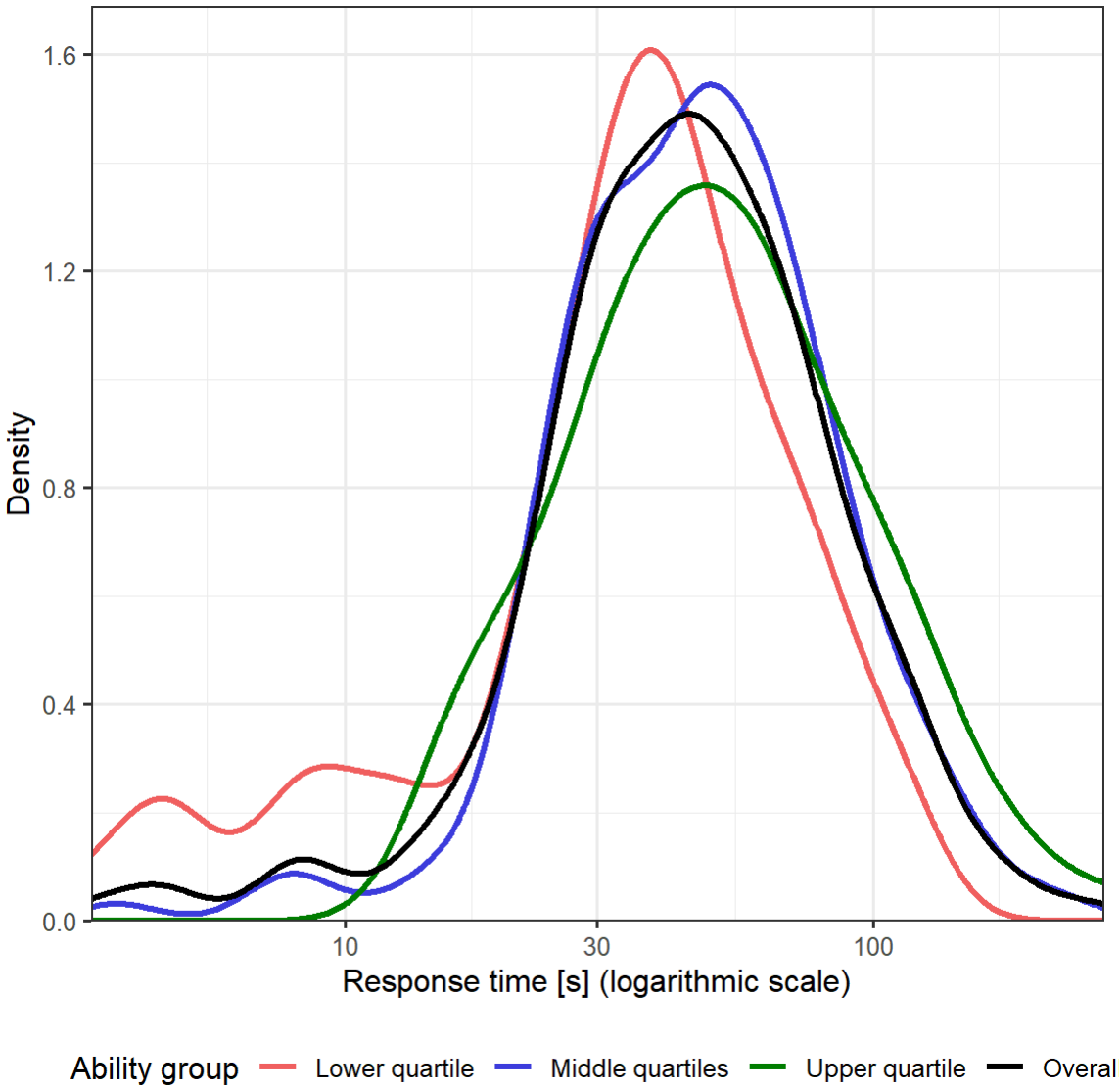
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
10	7	4	2	13	15.2%	5.3%	8.0%	8.9%
26	8	24	8	40	17.4%	32.0%	32.0%	27.4%
31	31	47	15	93	67.4%	62.7%	60.0%	63.7%
Overall	46	75	25	146	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



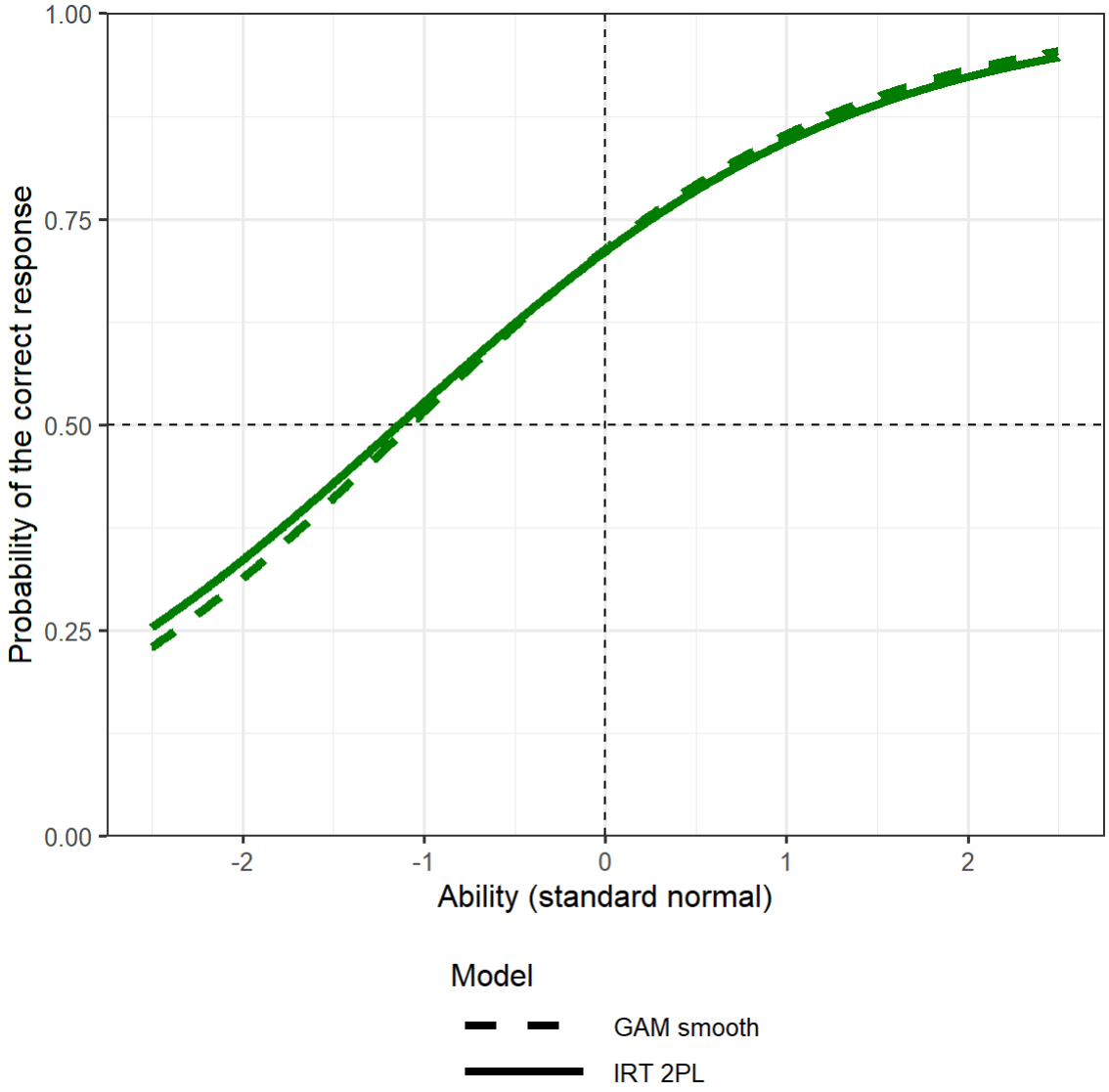
10.1.19.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.3	25.5	36.3	51.0	114.2	3.5	0.8
Middle quartiles	3.7	31.3	46.9	65.9	231.8	3.8	0.6
Upper quartile	15.8	32.4	50.4	76.6	273.1	3.9	0.6
Overall	3.3	30.3	45.0	65.6	273.1	3.8	0.7

10.1.20 M890022 (LLM)

10.1.20.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	68.97%	319	LLM complex setup	MC	0.35	0.27	0.79	−1.14	0.03

10.1.20.2 Gender DIF:

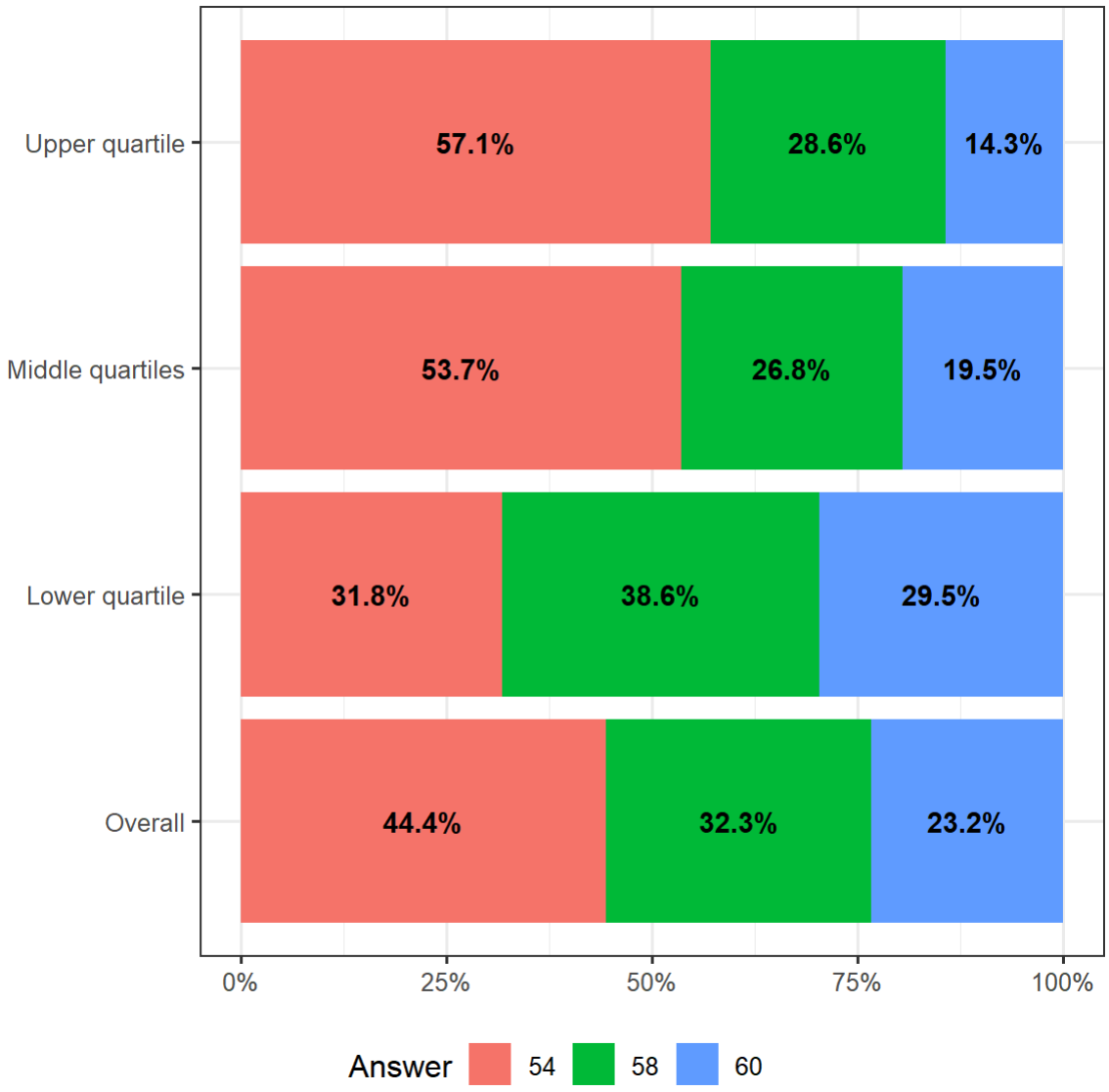
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.26	0.877	1.000	0.001	N/A

10.1.20.3 Frequency of answers:

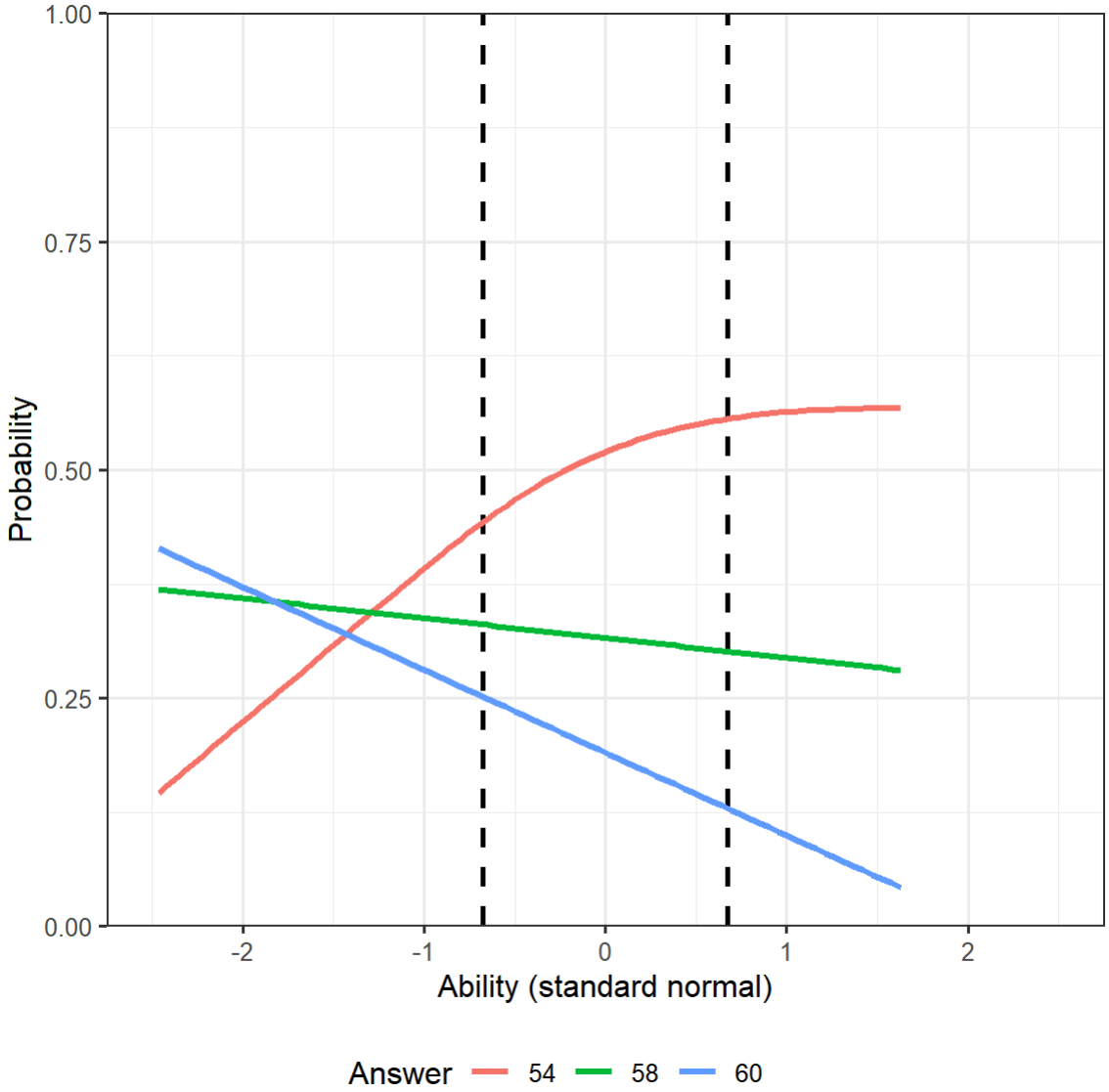
Answer	Frequency	Percent
54	44	13.8%
56 (correct)	220	69.0%
58	32	10.0%
60	23	7.2%
Overall	319	100.0%

10.1.20.4 Distractors functioning:

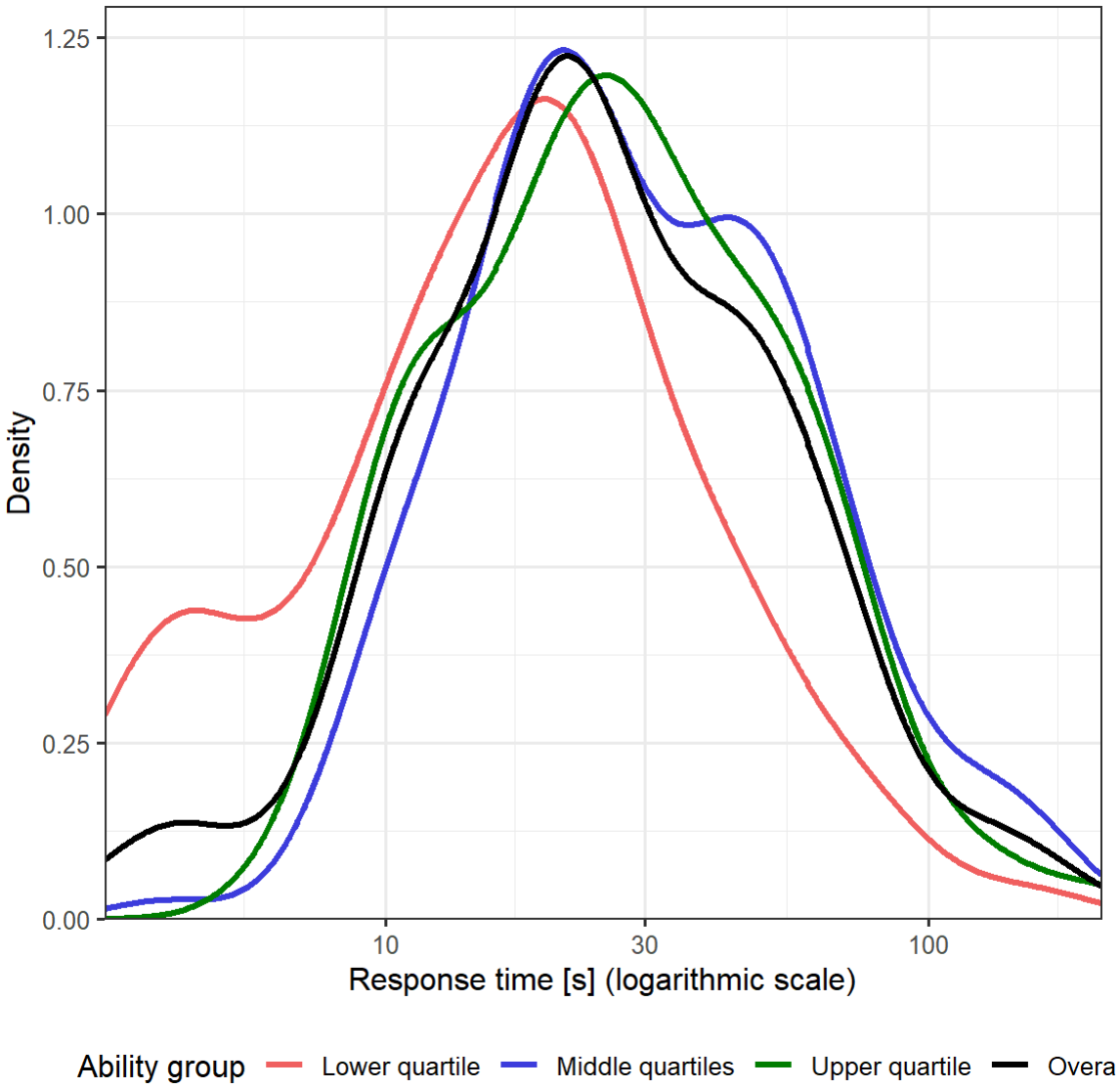
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
54	14	22	8	44	31.8%	53.7%	57.1%	44.4%
58	17	11	4	32	38.6%	26.8%	28.6%	32.3%
60	13	8	2	23	29.5%	19.5%	14.3%	23.2%
Overall	44	41	14	99	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



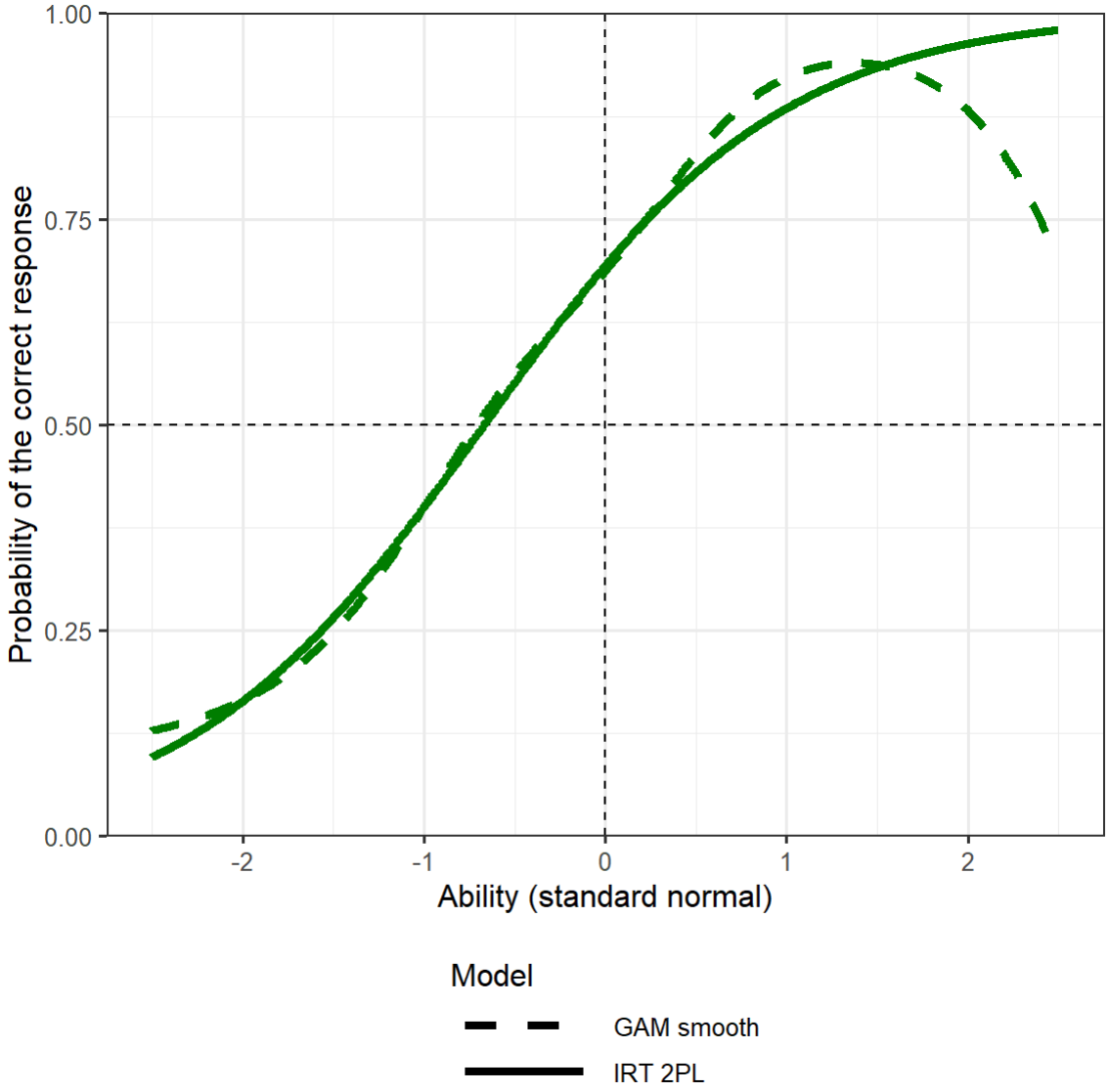
10.1.20.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	9.9	18.1	27.1	154.0	2.8	0.8
Middle quartiles	3.9	18.4	26.1	48.3	180.8	3.4	0.7
Upper quartile	6.5	16.9	26.4	45.2	208.2	3.3	0.7
Overall	3.0	15.1	24.1	43.8	208.2	3.2	0.8

10.1.21 M061052 (TIMSS)

10.1.21.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	66.33%	294	TIMSS	MC	0.45	0.45	1.22	−0.67	0.019

10.1.21.2 Gender DIF:

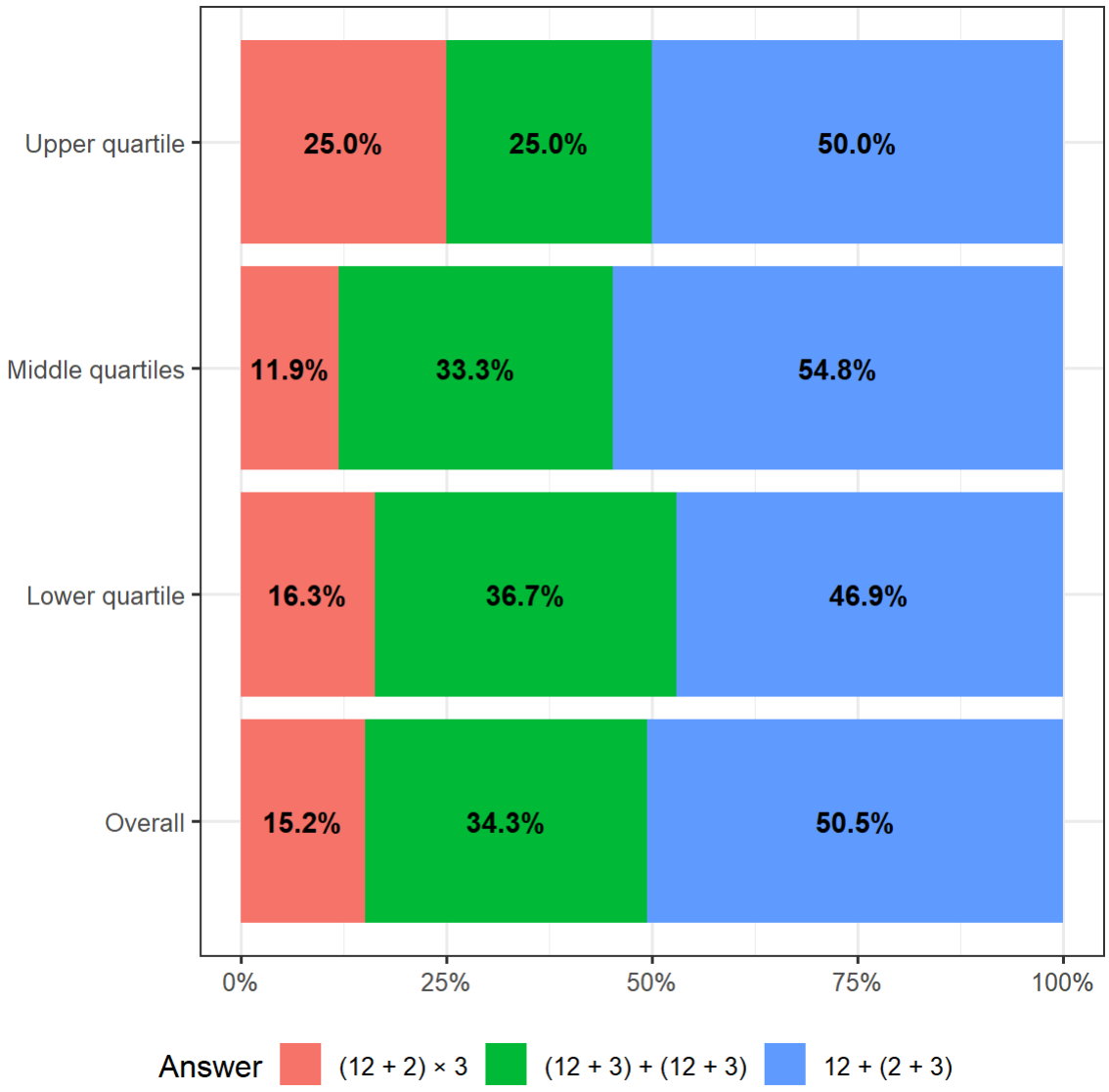
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.95	0.377	1.000	0.006	N/A

10.1.21.3 Frequency of answers:

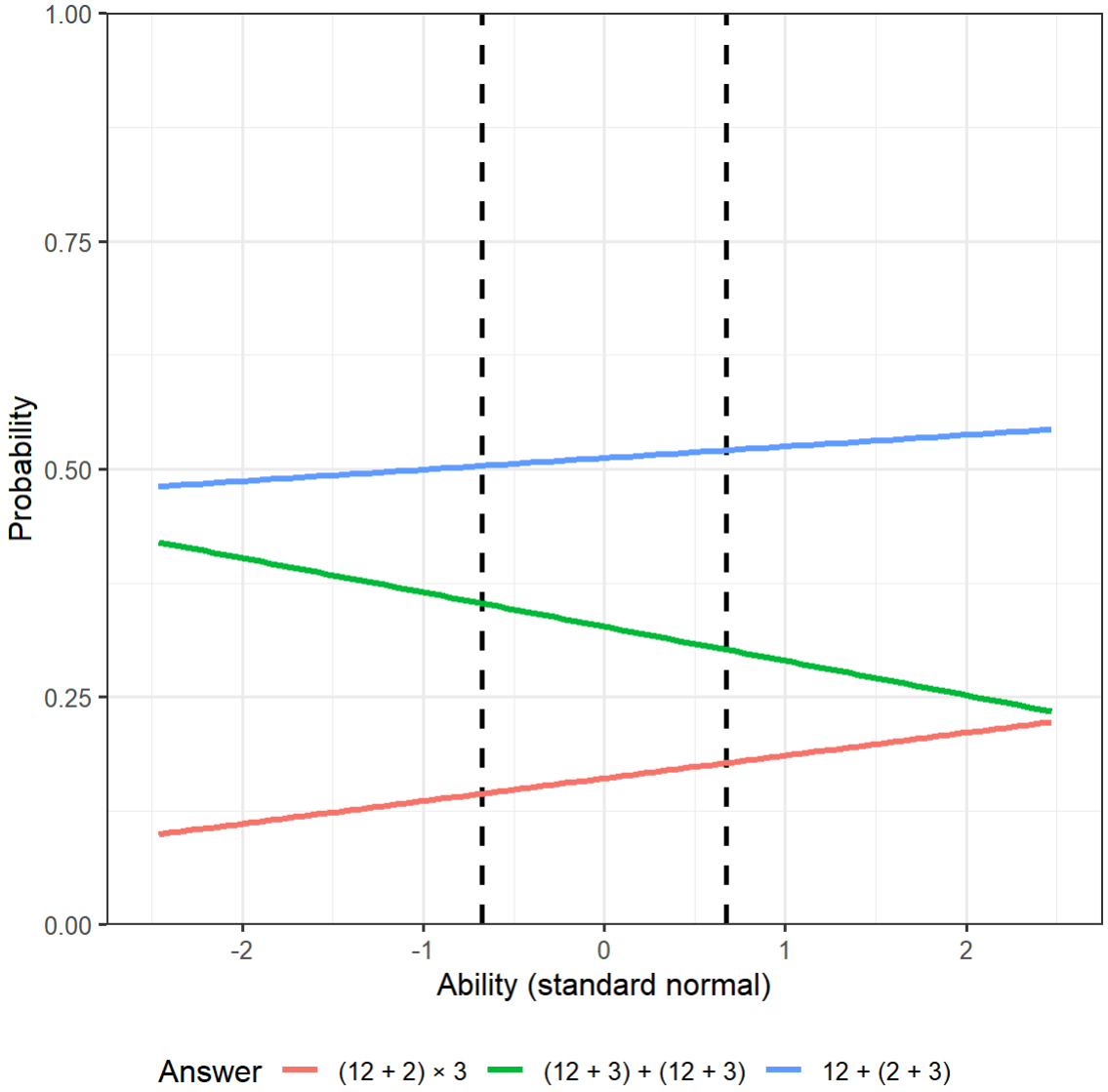
Answer	Frequency	Percent
(12 + 2) × 3	15	5.1%
(12 + 3) + (12 + 3)	34	11.6%
12 + (2 + 3)	50	17.0%
12 + (2 × 3) (correct)	195	66.3%
Overall	294	100.0%

10.1.21.4 Distractors functioning:

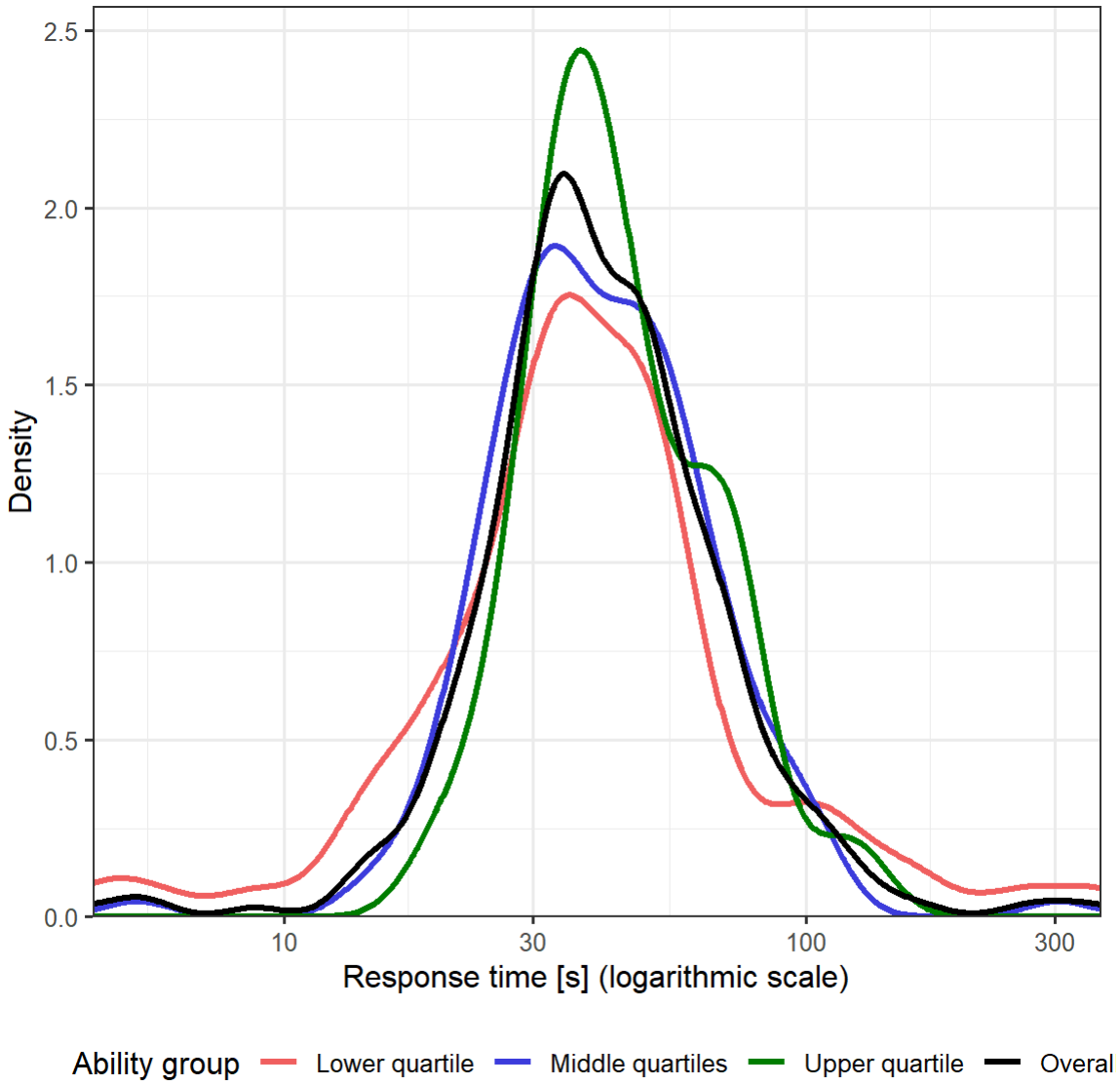
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
(12 + 2) × 3	8	5	2	15	16.3%	11.9%	25.0%	15.2%
(12 + 3) + (12 + 3)	18	14	2	34	36.7%	33.3%	25.0%	34.3%
12 + (2 + 3)	23	23	4	50	46.9%	54.8%	50.0%	50.5%
Overall	49	42	8	99	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



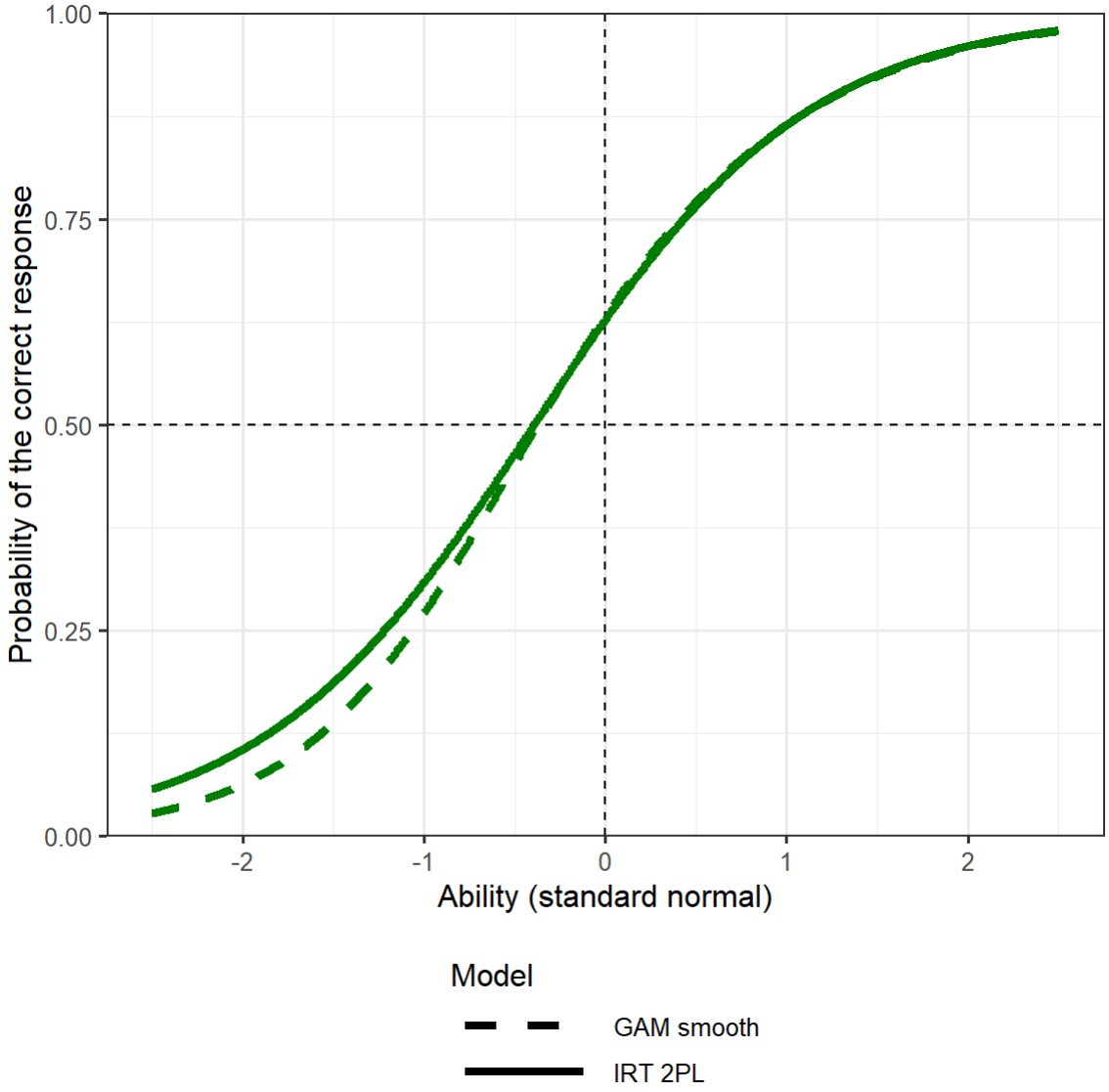
10.1.21.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.3	27.7	36.3	50.9	366.8	3.6	0.7
Middle quartiles	5.2	29.4	39.5	54.0	307.4	3.7	0.5
Upper quartile	18.8	33.6	41.1	58.5	127.3	3.8	0.4
Overall	4.3	30.7	38.9	53.9	366.8	3.7	0.5

10.1.22 M061029 (TIMSS)

10.1.22.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	58.88%	304	TIMSS	MC	0.52	0.49	1.33	−0.39	0.009

10.1.22.2 Gender DIF:

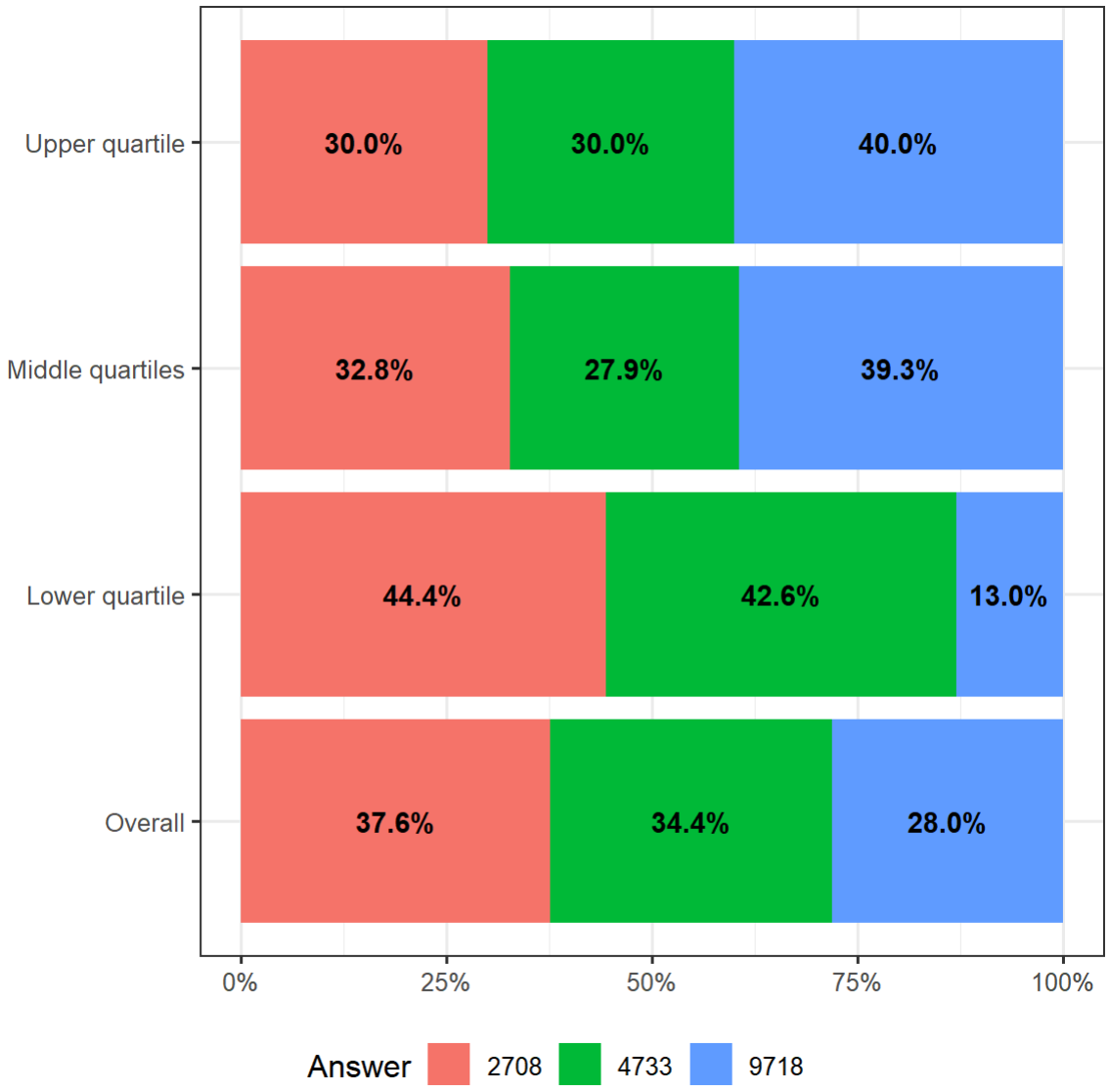
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.04	0.981	1.000	0.000	N/A

10.1.22.3 Frequency of answers:

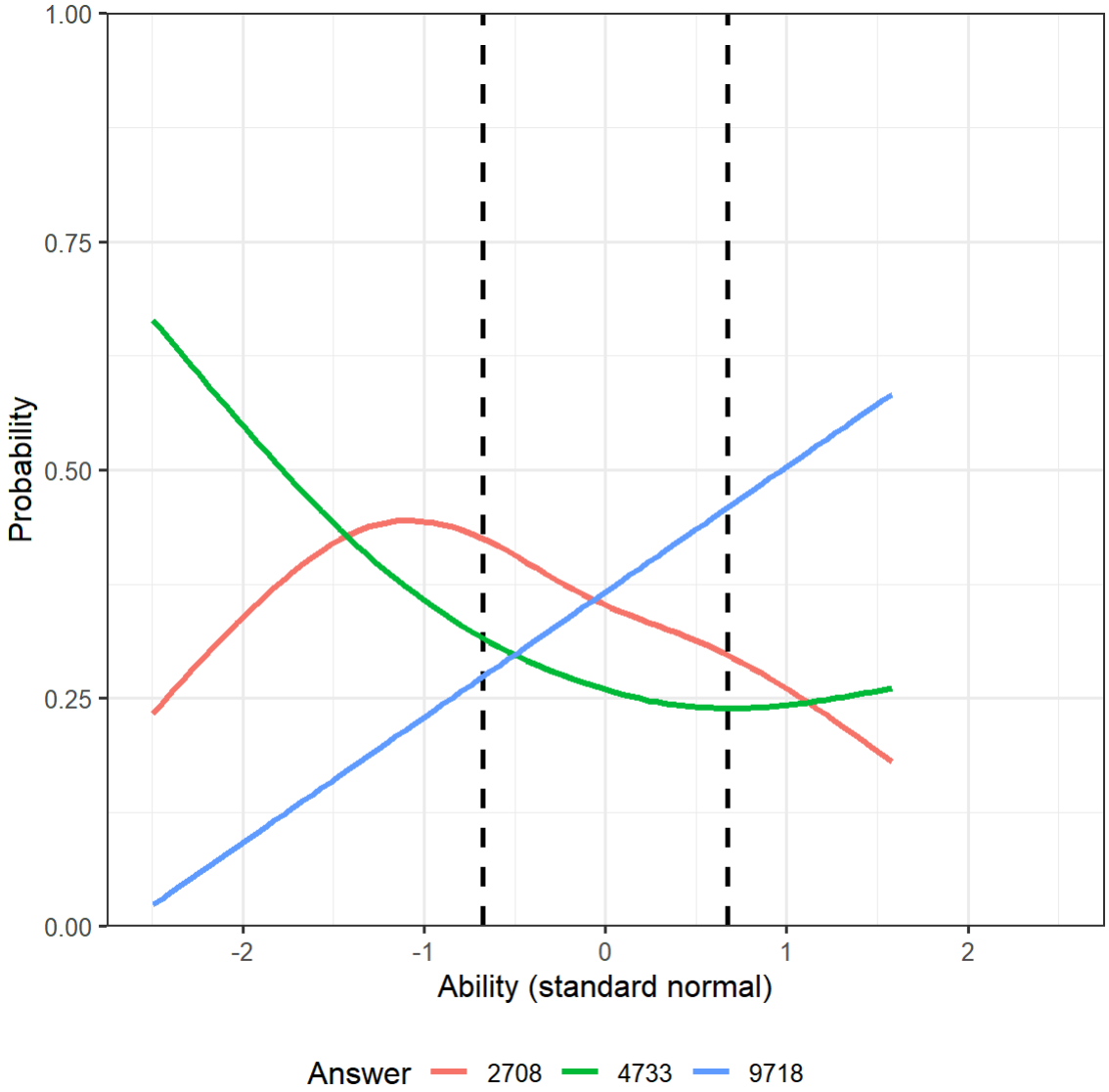
Answer	Frequency	Percent
2708	47	15.5%
4733	43	14.1%
8726 (correct)	179	58.9%
9718	35	11.5%
Overall	304	100.0%

10.1.22.4 Distractors functioning:

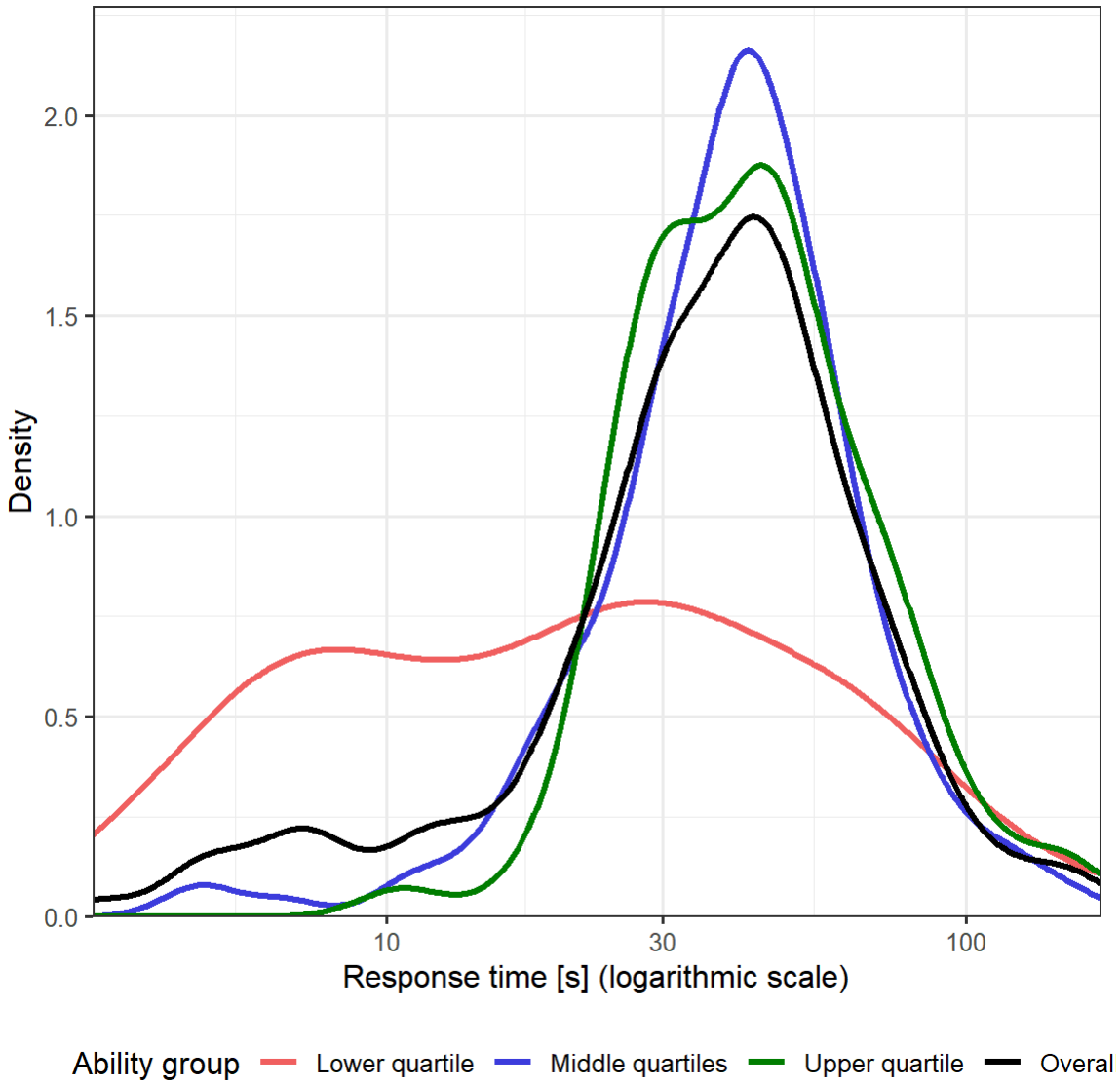
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
2708	24	20	3	47	44.4%	32.8%	30.0%	37.6%
4733	23	17	3	43	42.6%	27.9%	30.0%	34.4%
9718	7	24	4	35	13.0%	39.3%	40.0%	28.0%
Overall	54	61	10	125	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



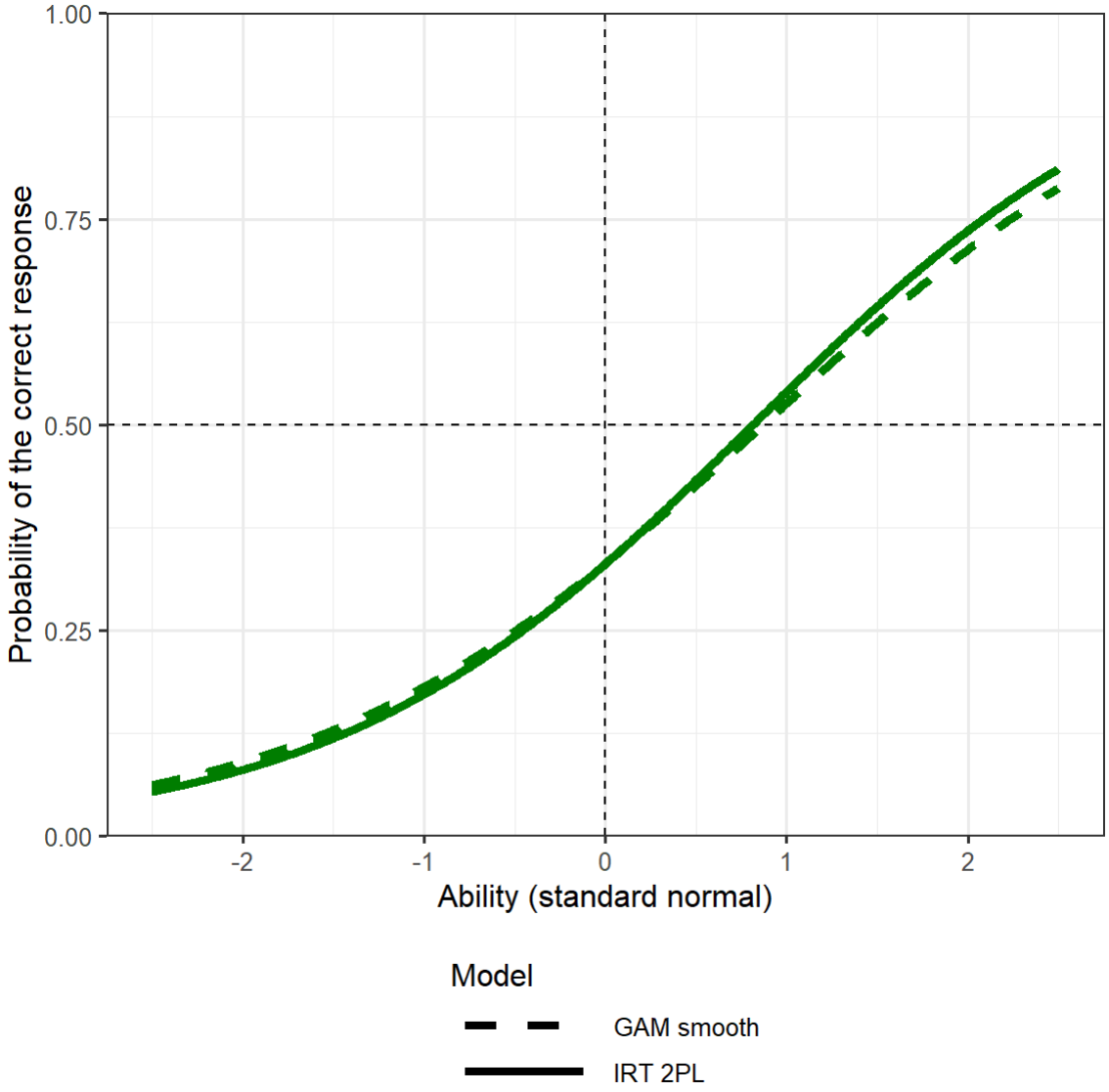
10.1.22.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	8.1	21.4	42.9	170.8	3.0	1.0
Middle quartiles	4.5	29.9	40.9	53.4	158.5	3.7	0.5
Upper quartile	10.7	30.0	42.8	54.9	152.8	3.8	0.5
Overall	3.1	26.2	38.9	52.9	170.8	3.5	0.7

10.1.23 M041125 (TIMSS)

10.1.23.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	34.43%	306	TIMSS	MC	0.34	0.37	0.86	0.82	0.000

10.1.23.2 Gender DIF:

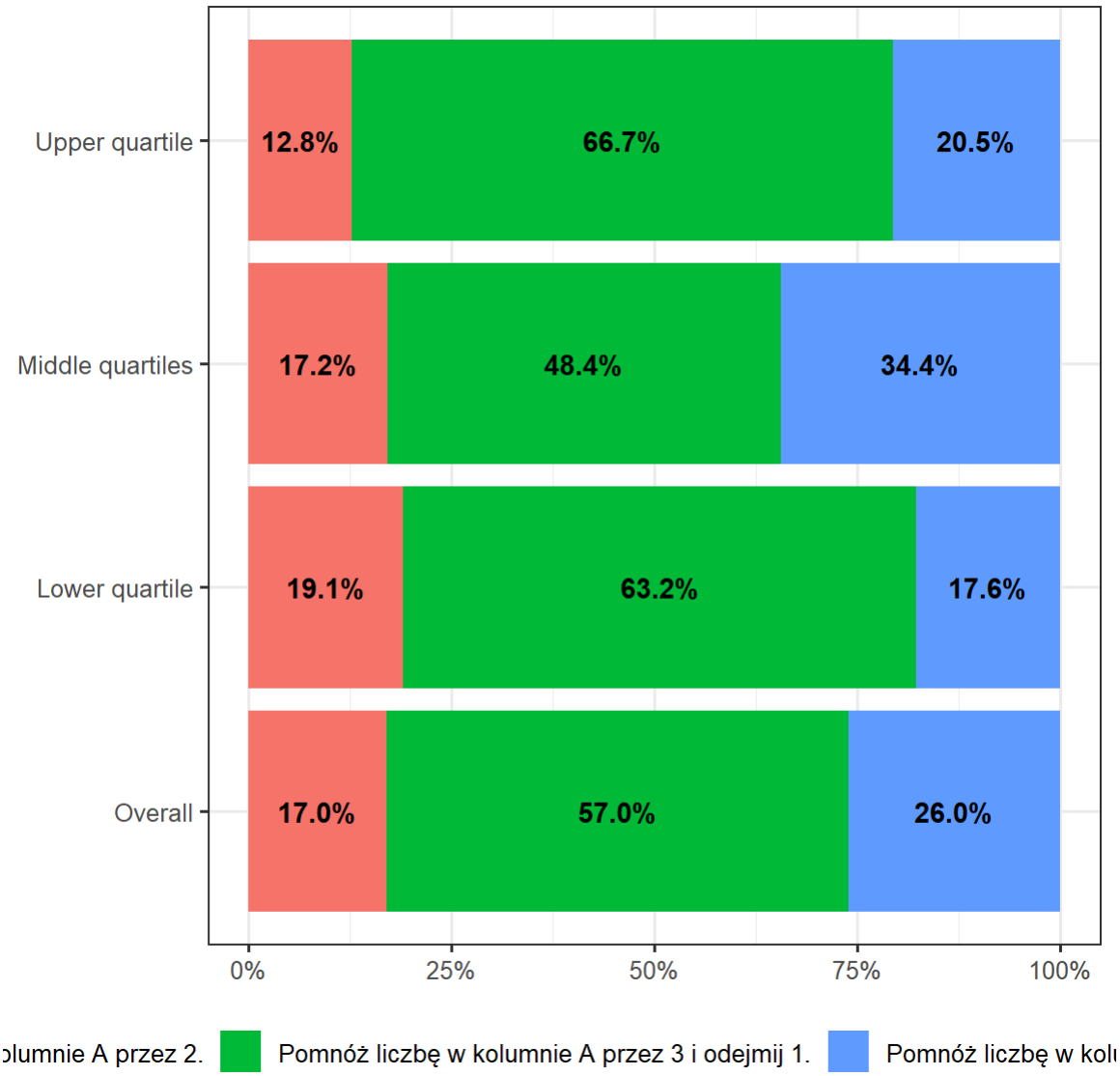
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.06	0.969	1.000	0.000	N/A

10.1.23.3 Frequency of answers:

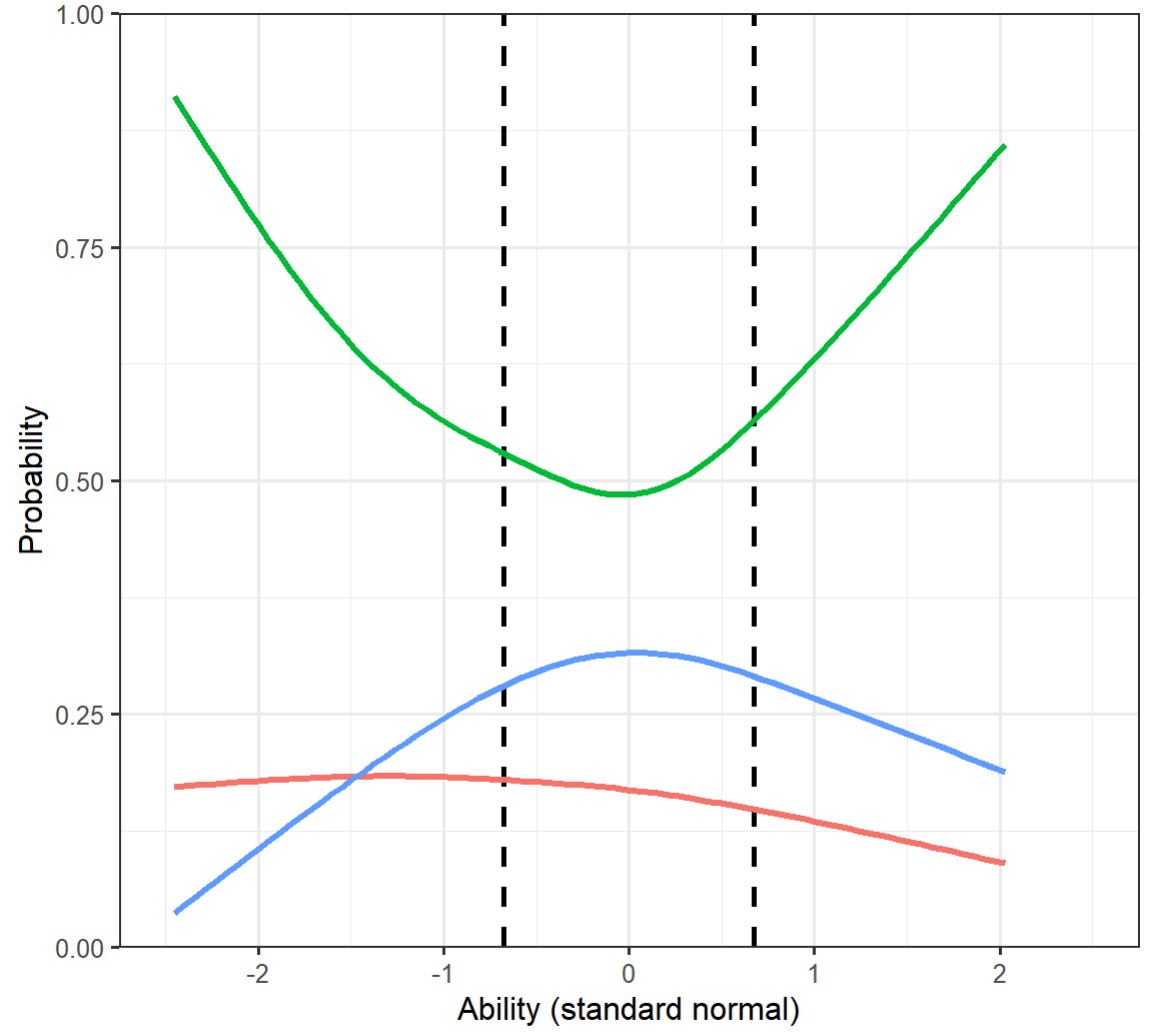
Answer	Frequency	Percent
Pomnóż liczbę w kolumnie A przez 2.	34	11.1%
Pomnóż liczbę w kolumnie A przez 3 i odejmij 1.	114	37.4%
Pomnóż liczbę w kolumnie A przez nią samą i dodaj 1. (correct)	105	34.4%
Pomnóż liczbę w kolumnie A przez nią samą i odejmij 1.	52	17.0%
Overall	305	100.0%

10.1.23.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Pomnóż liczbę w kolumnie A przez 2.	13	16	5	34	19.1%	17.2%	12.8%	17.0%
Pomnóż liczbę w kolumnie A przez 3 i odejmij 1.	43	45	26	114	63.2%	48.4%	66.7%	57.0%
Pomnóż liczbę w kolumnie A przez nią samą i odejmij 1.	12	32	8	52	17.6%	34.4%	20.5%	26.0%
Overall	68	93	39	200	100.0%	100.0%	100.0%	100.0%

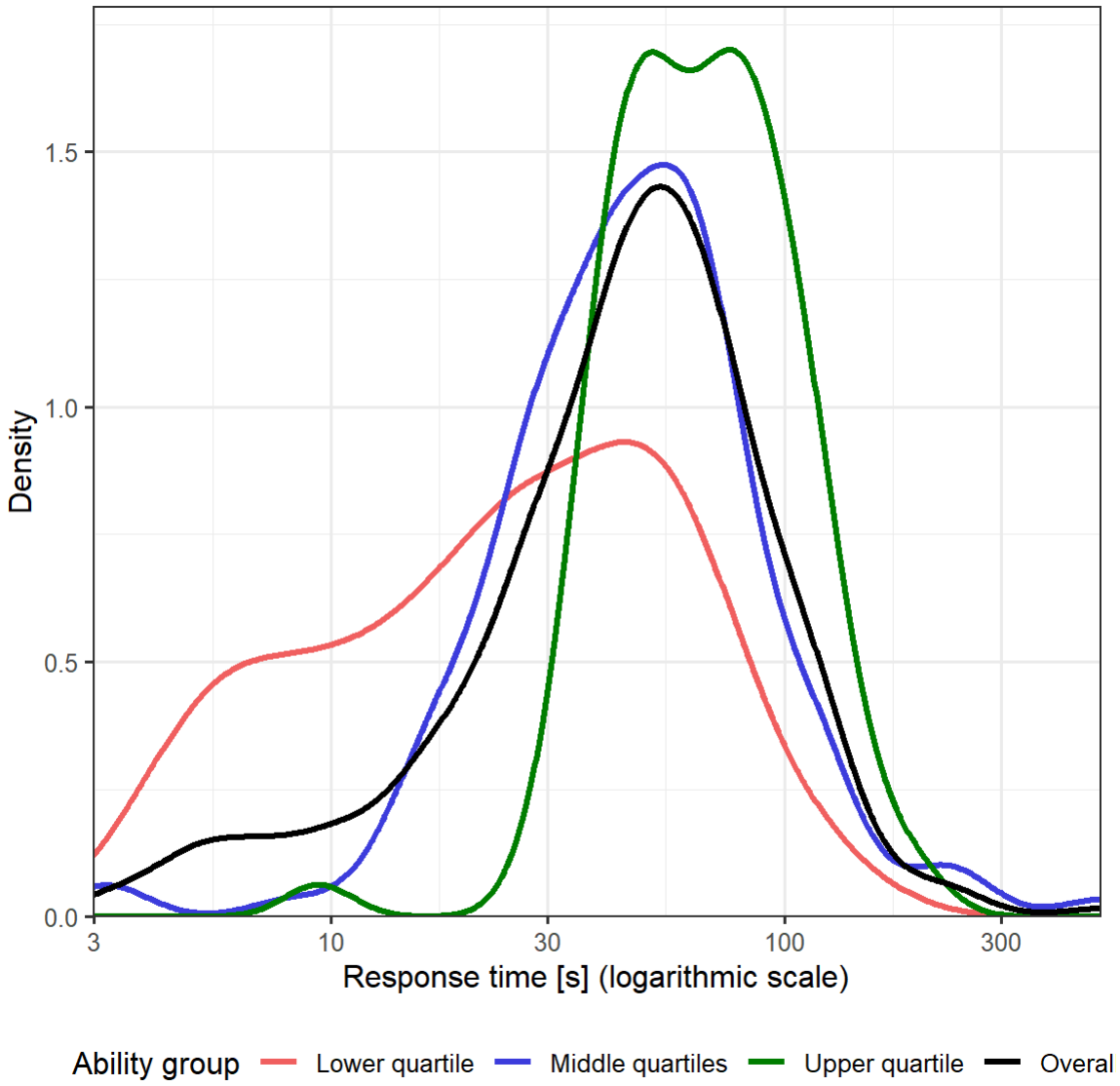



```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



olumnie A przez 2. Pomnóż liczbę w kolumnie A przez 3 i odejmij 1. Pomnóż liczbę w kol

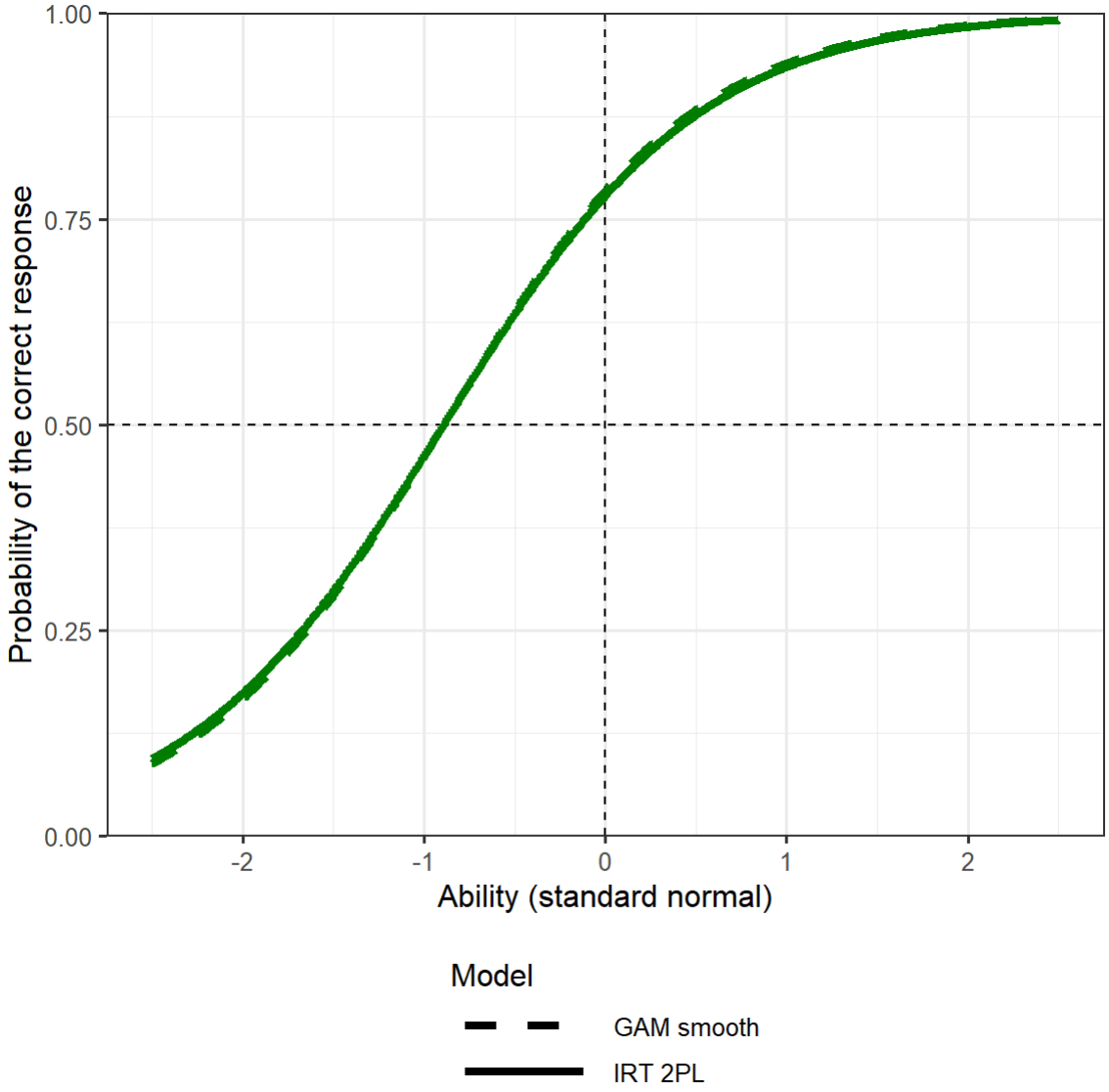
10.1.23.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.6	11.2	27.3	51.5	134.1	3.2	0.9
Middle quartiles	3.0	29.8	45.7	66.0	497.0	3.8	0.7
Upper quartile	9.4	47.0	66.5	91.8	199.5	4.2	0.5
Overall	3.0	28.5	47.2	70.2	497.0	3.7	0.8

10.1.24 MP71219 (TIMSS)

10.1.24.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	72.48%	299	TIMSS	MC	0.49	0.51	1.41	−0.89	0.000

10.1.24.2 Gender DIF:

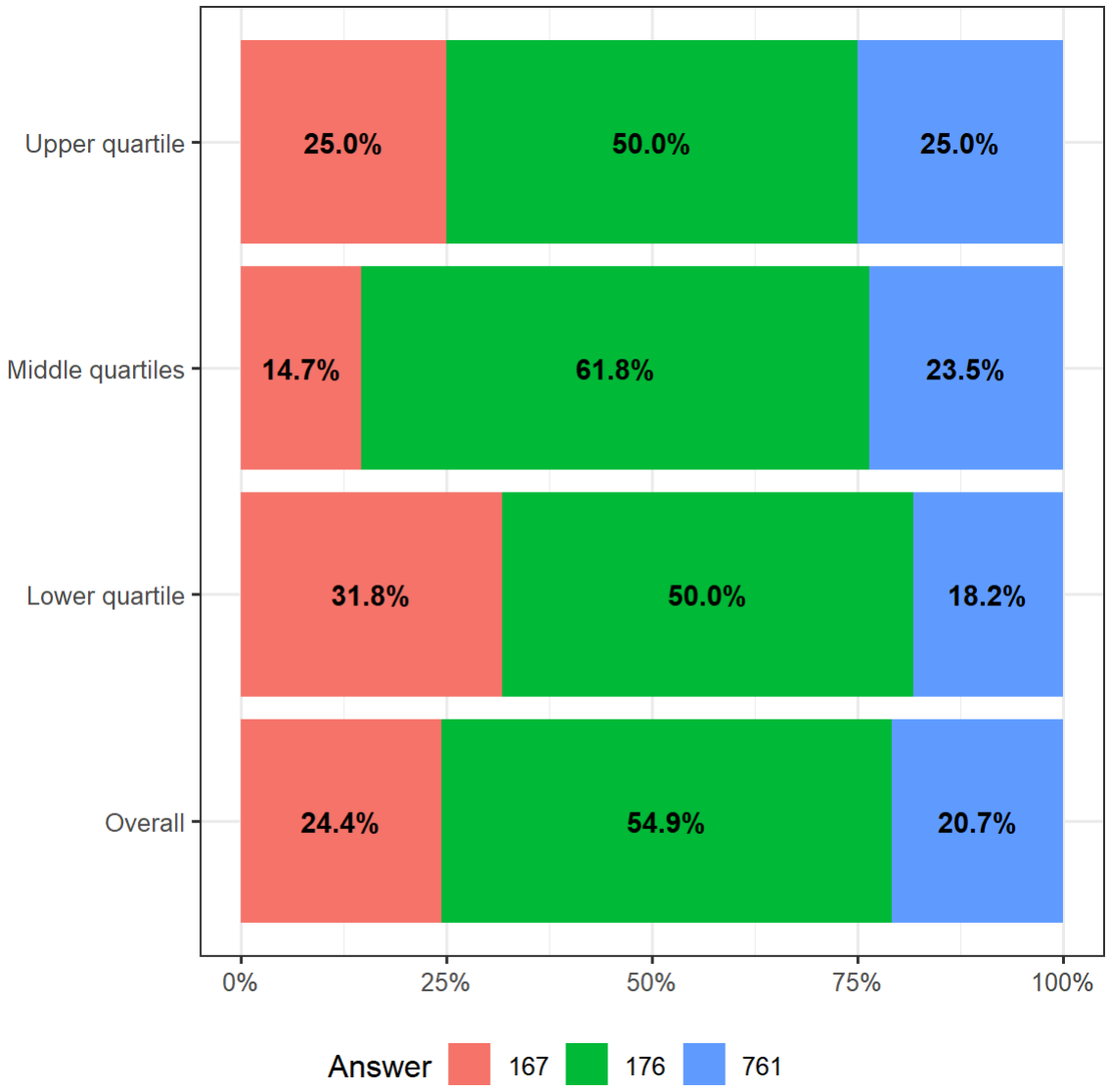
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.47	0.792	1.000	0.001	N/A

10.1.24.3 Frequency of answers:

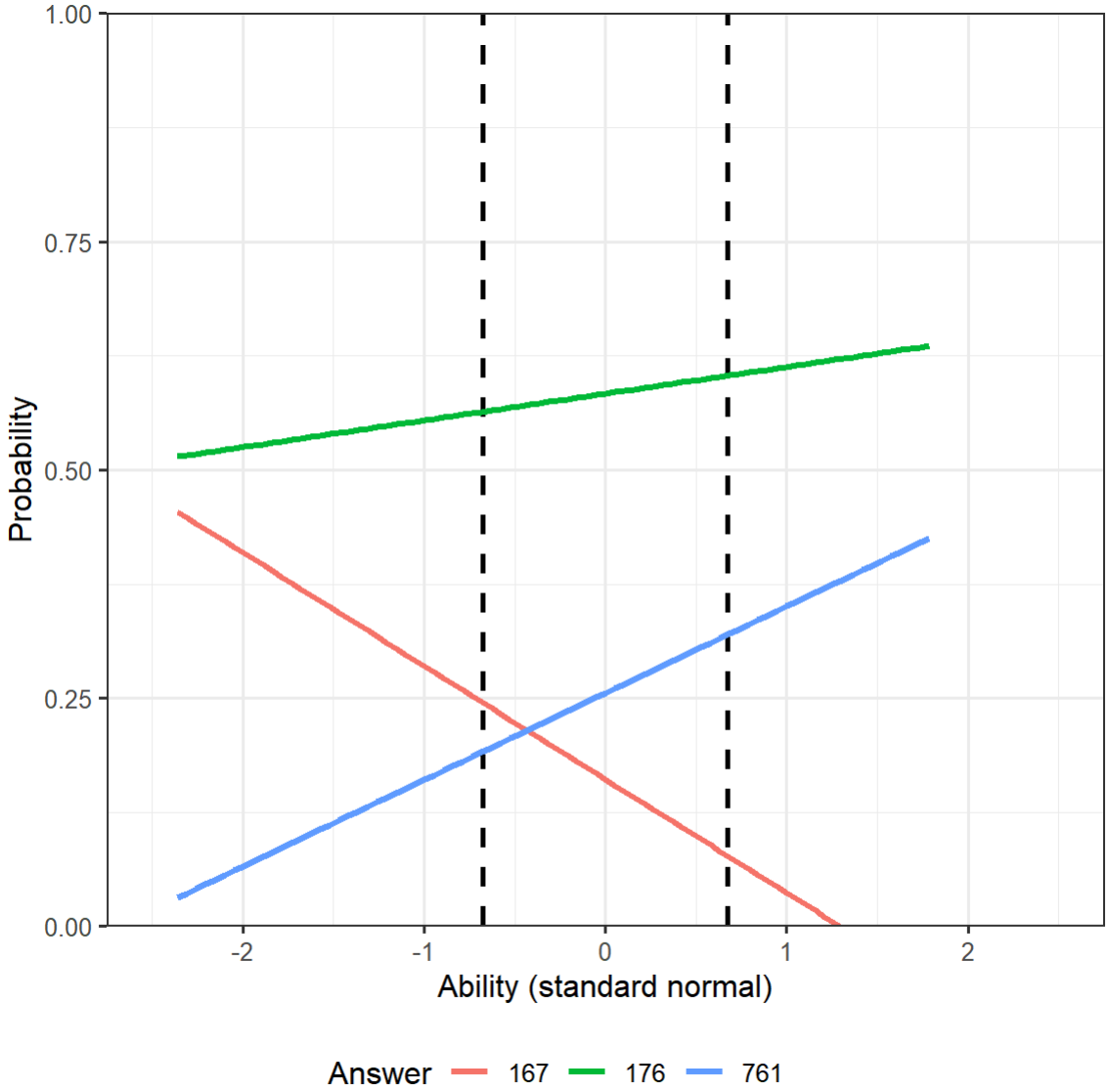
Answer	Frequency	Percent
167	20	6.7%
176	45	15.1%
716 (correct)	216	72.5%
761	17	5.7%
Overall	298	100.0%

10.1.24.4 Distractors functioning:

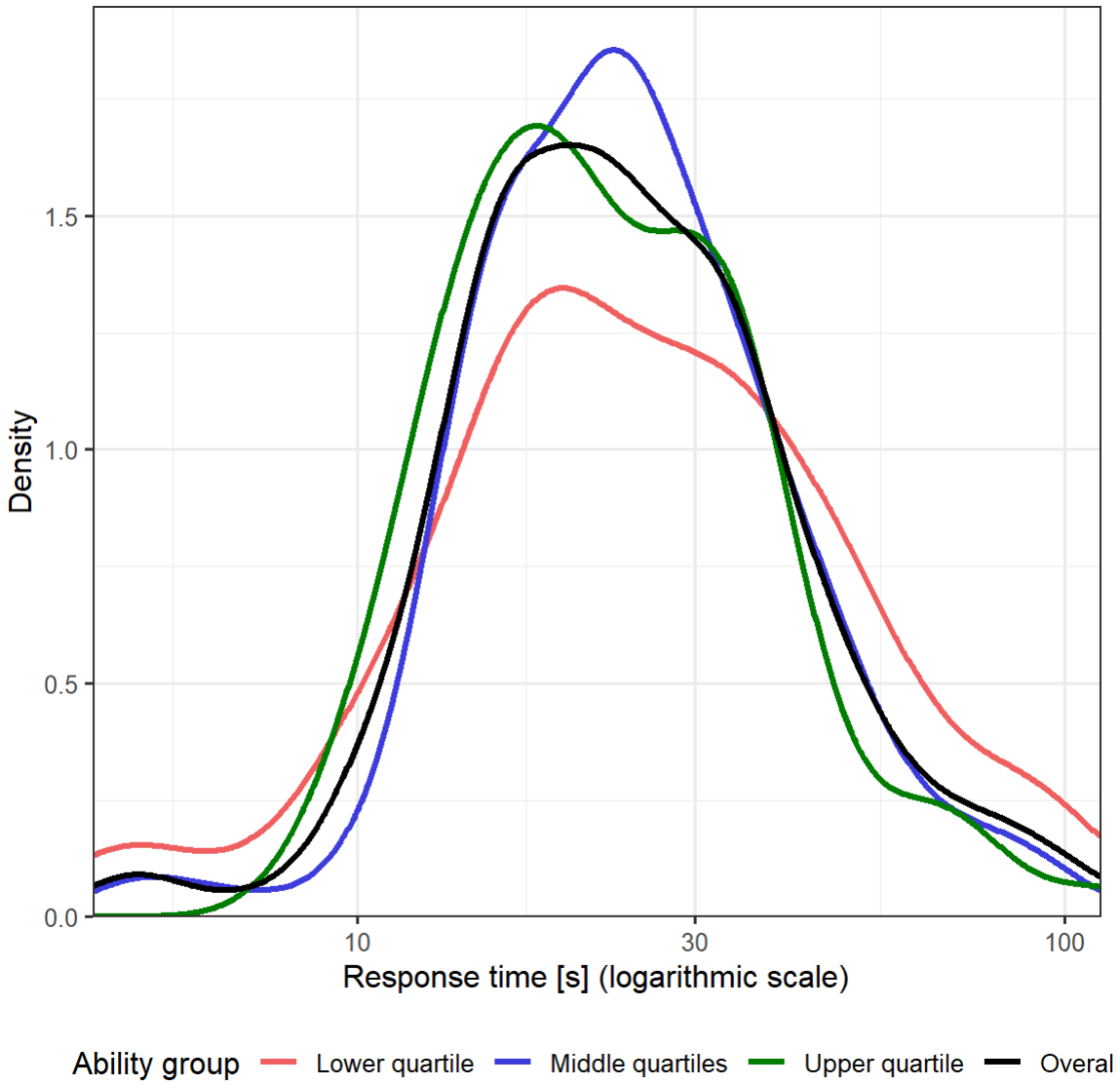
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
167	14	5	1	20	31.8%	14.7%	25.0%	24.4%
176	22	21	2	45	50.0%	61.8%	50.0%	54.9%
761	8	8	1	17	18.2%	23.5%	25.0%	20.7%
Overall	44	34	4	82	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



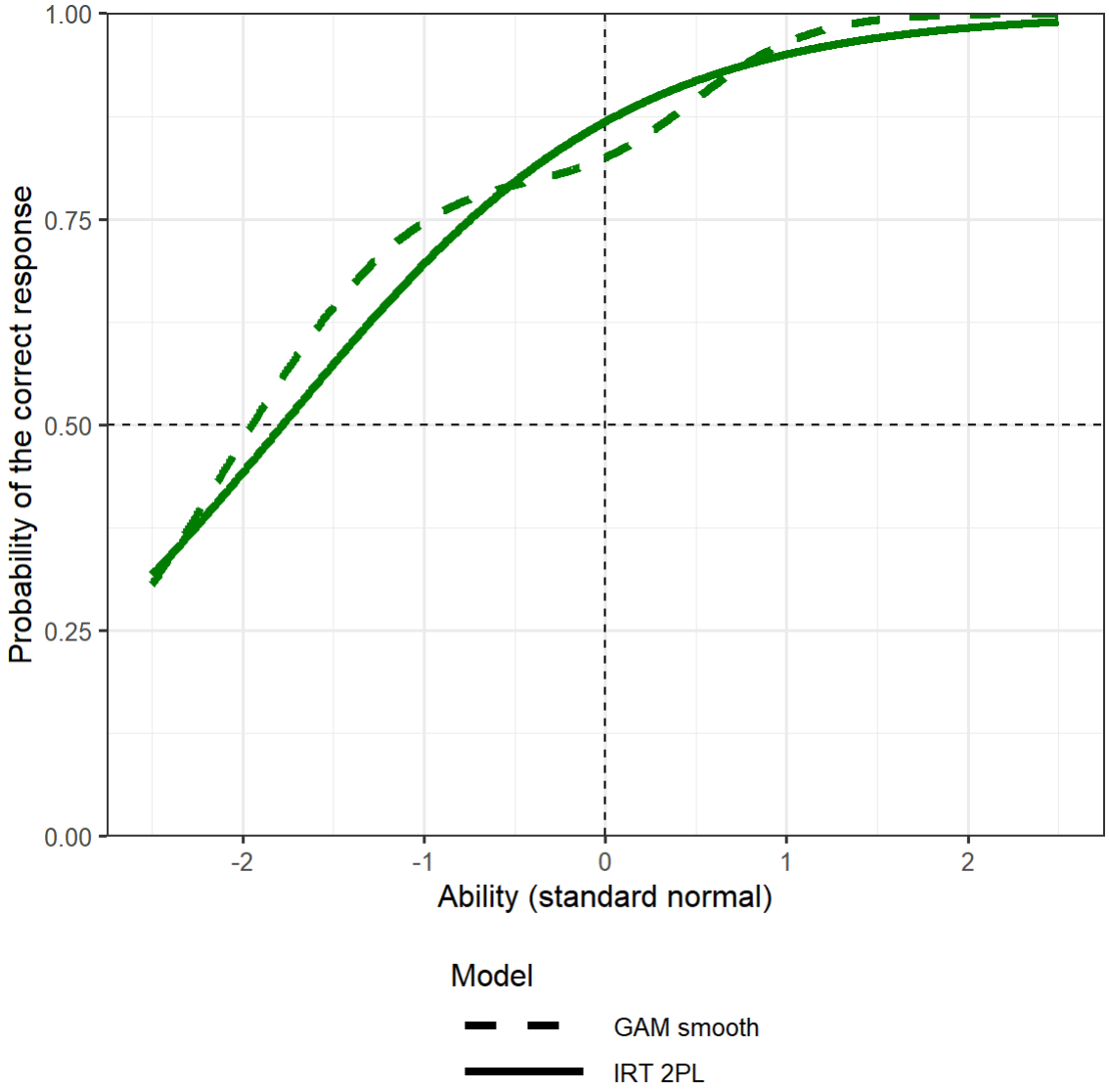
10.1.24.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.2	16.7	24.2	37.1	110.5	3.2	0.7
Middle quartiles	4.7	16.6	23.3	32.7	99.9	3.2	0.5
Upper quartile	8.7	15.5	21.2	32.4	112.2	3.1	0.5
Overall	4.2	16.5	23.1	33.6	112.2	3.2	0.6

10.1.25 M041298 (TIMSS)

10.1.25.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	83.22%	299	TIMSS	MC	0.34	0.39	1.06	−1.78	0.000

10.1.25.2 Gender DIF:

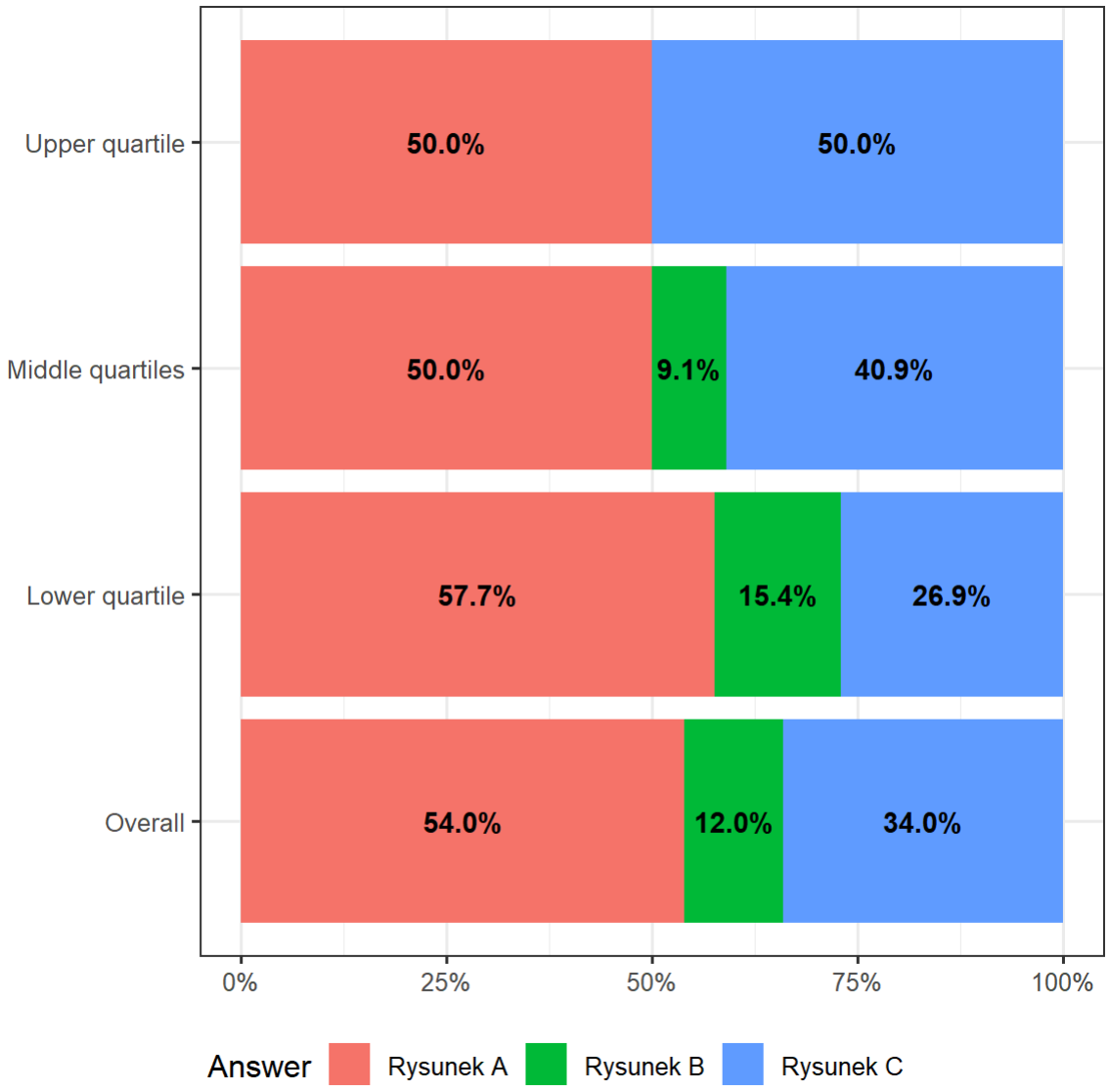
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.97	0.616	1.000	0.004	N/A

10.1.25.3 Frequency of answers:

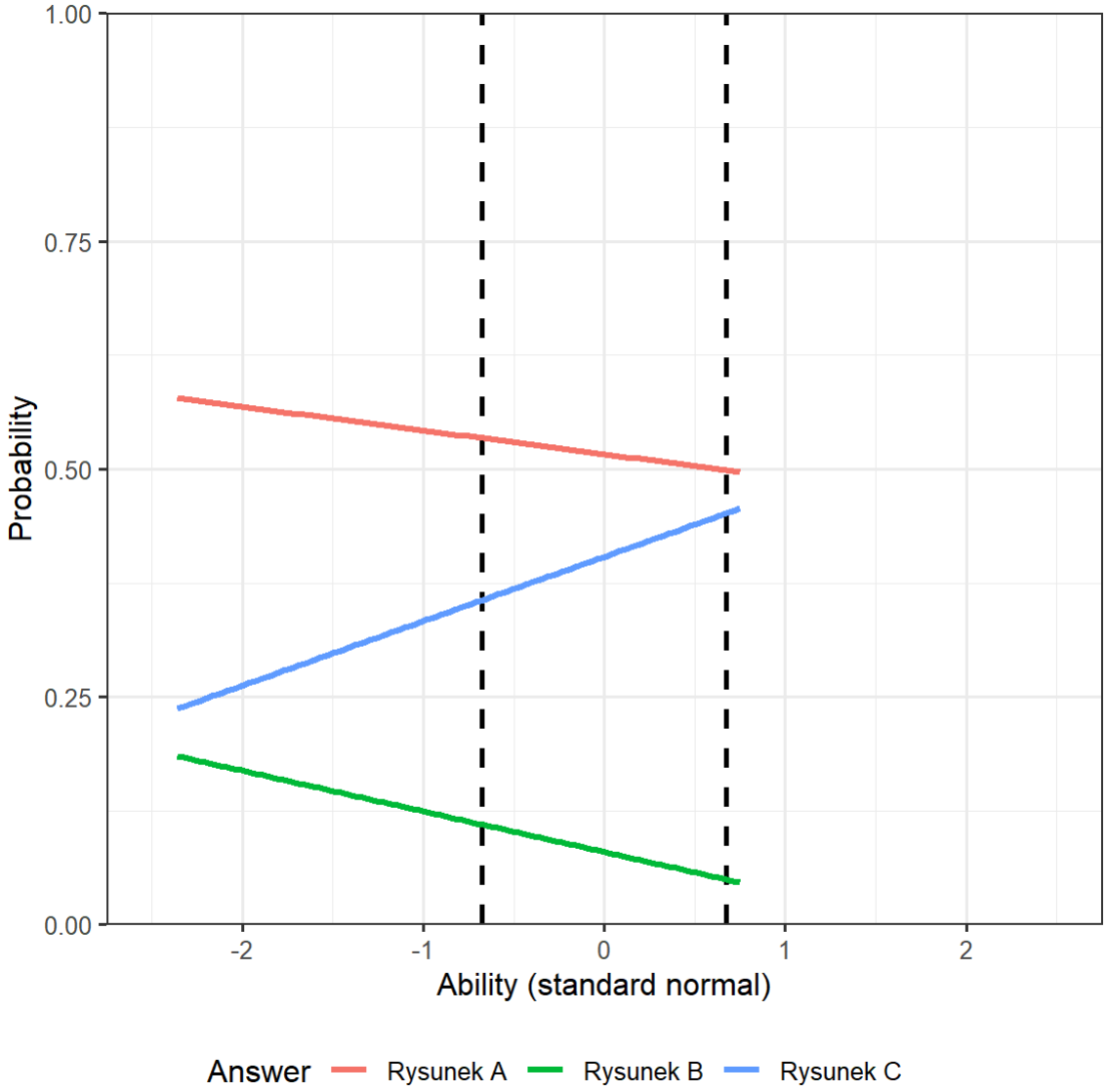
Answer	Frequency	Percent
Rysunek A	27	9.1%
Rysunek B	6	2.0%
Rysunek C	17	5.7%
Rysunek D (correct)	248	83.2%
Overall	298	100.0%

10.1.25.4 Distractors functioning:

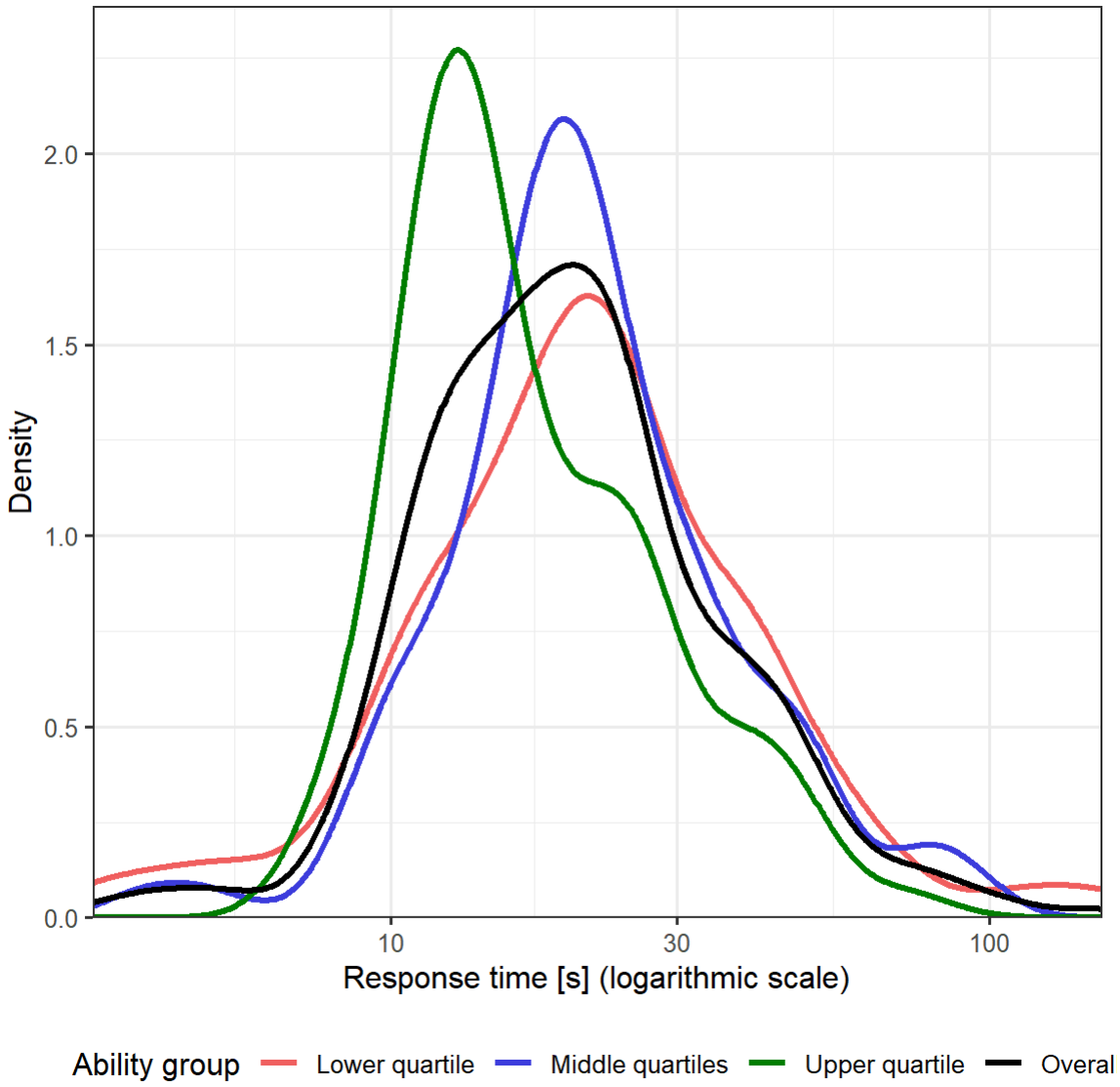
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Rysunek A	15	11	1	27	57.7%	50.0%	50.0%	54.0%
Rysunek B	4	2	0	6	15.4%	9.1%	0.0%	12.0%
Rysunek C	7	9	1	17	26.9%	40.9%	50.0%	34.0%
Overall	26	22	2	50	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



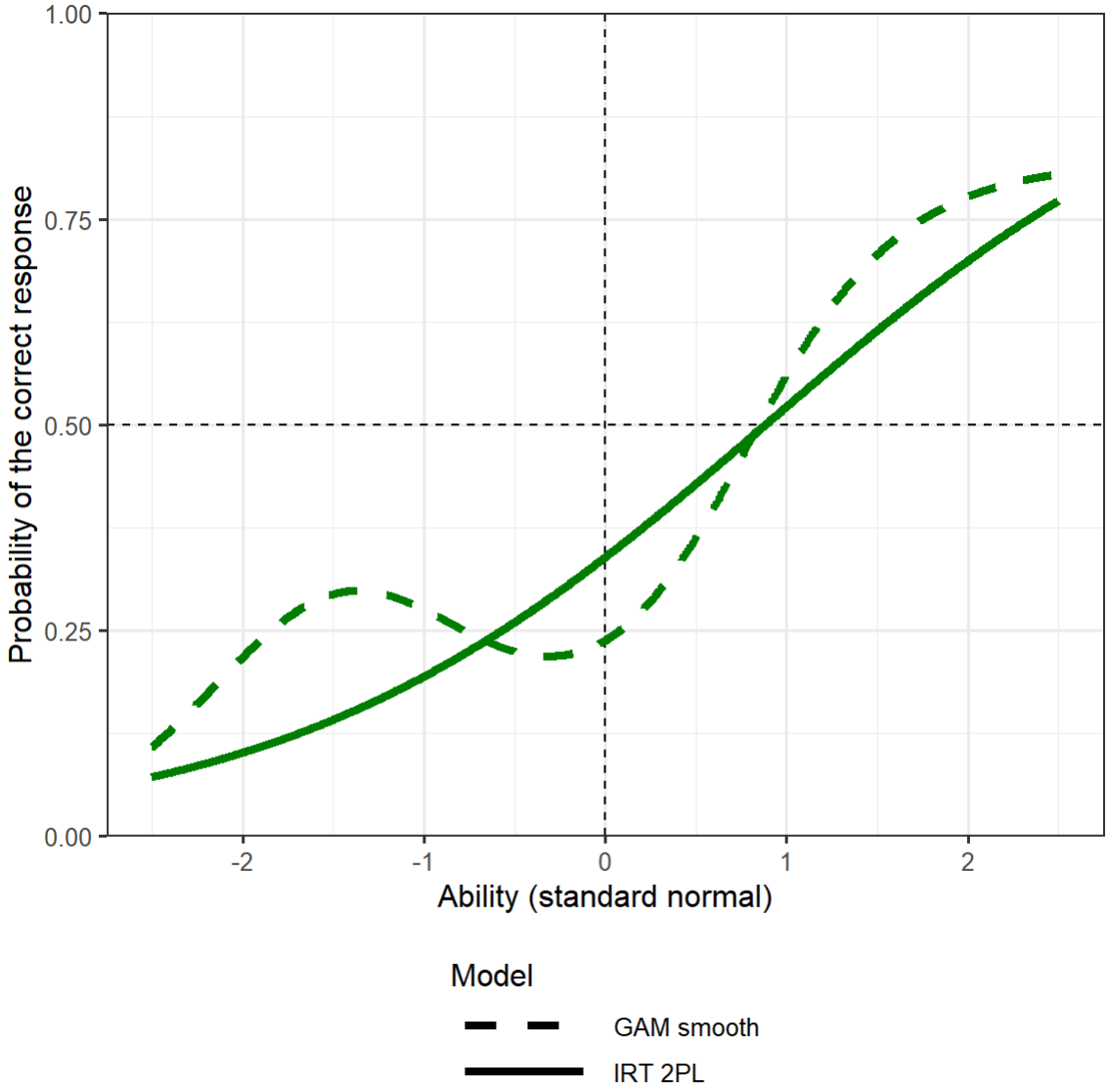
10.1.25.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.2	14.5	21.7	30.3	153.8	3.0	0.7
Middle quartiles	3.7	15.9	20.4	29.0	96.2	3.1	0.6
Upper quartile	7.6	11.9	14.4	23.7	72.9	2.8	0.5
Overall	3.2	13.5	19.2	27.0	153.8	3.0	0.6

10.1.26 M100026 (LLM)

10.1.26.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	35.86%	304	LLM simple setup	MC	0.29	0.25	0.76	0.88	0.035

10.1.26.2 Gender DIF:

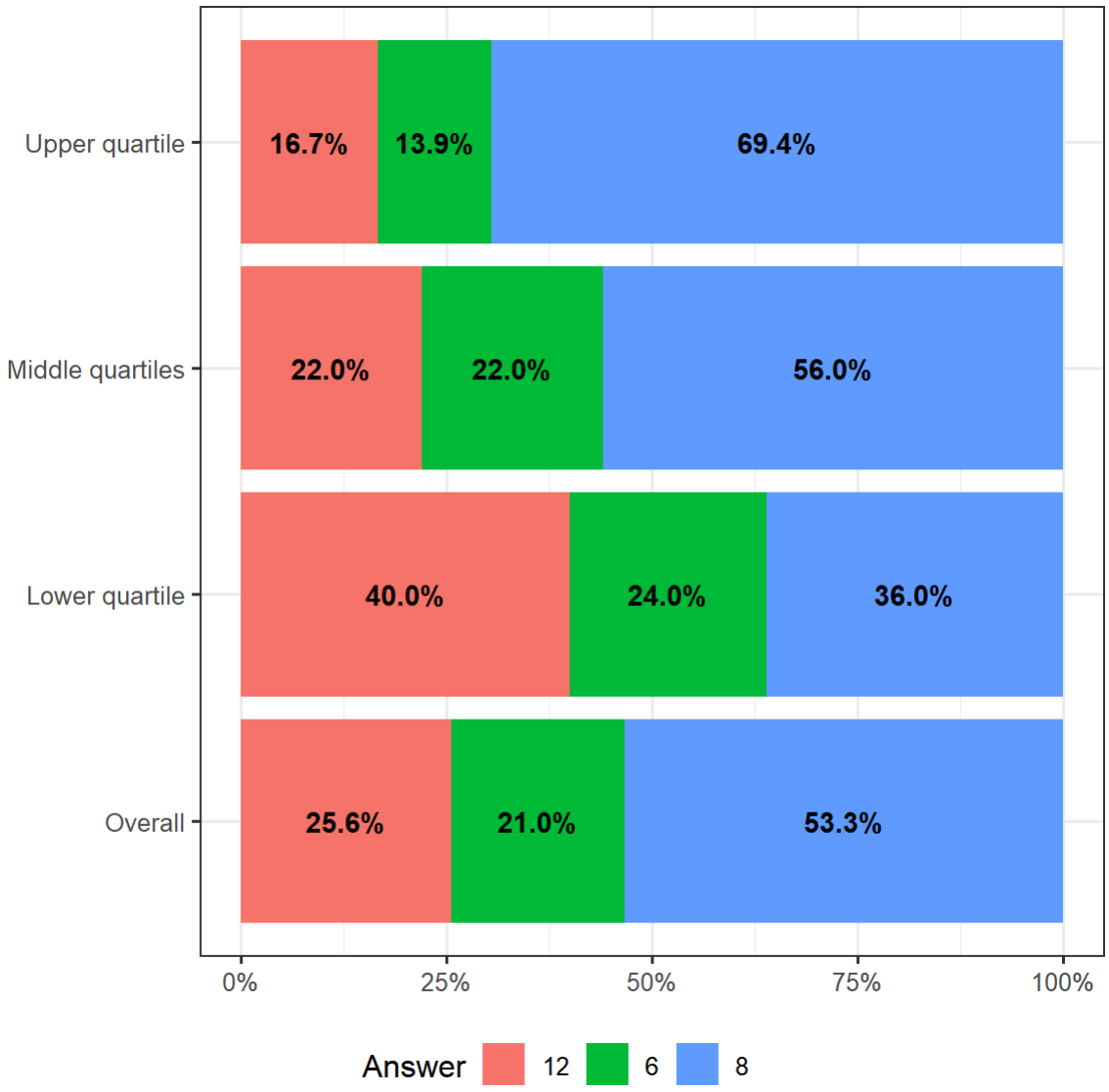
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.44	0.802	1.000	0.001	N/A

10.1.26.3 Frequency of answers:

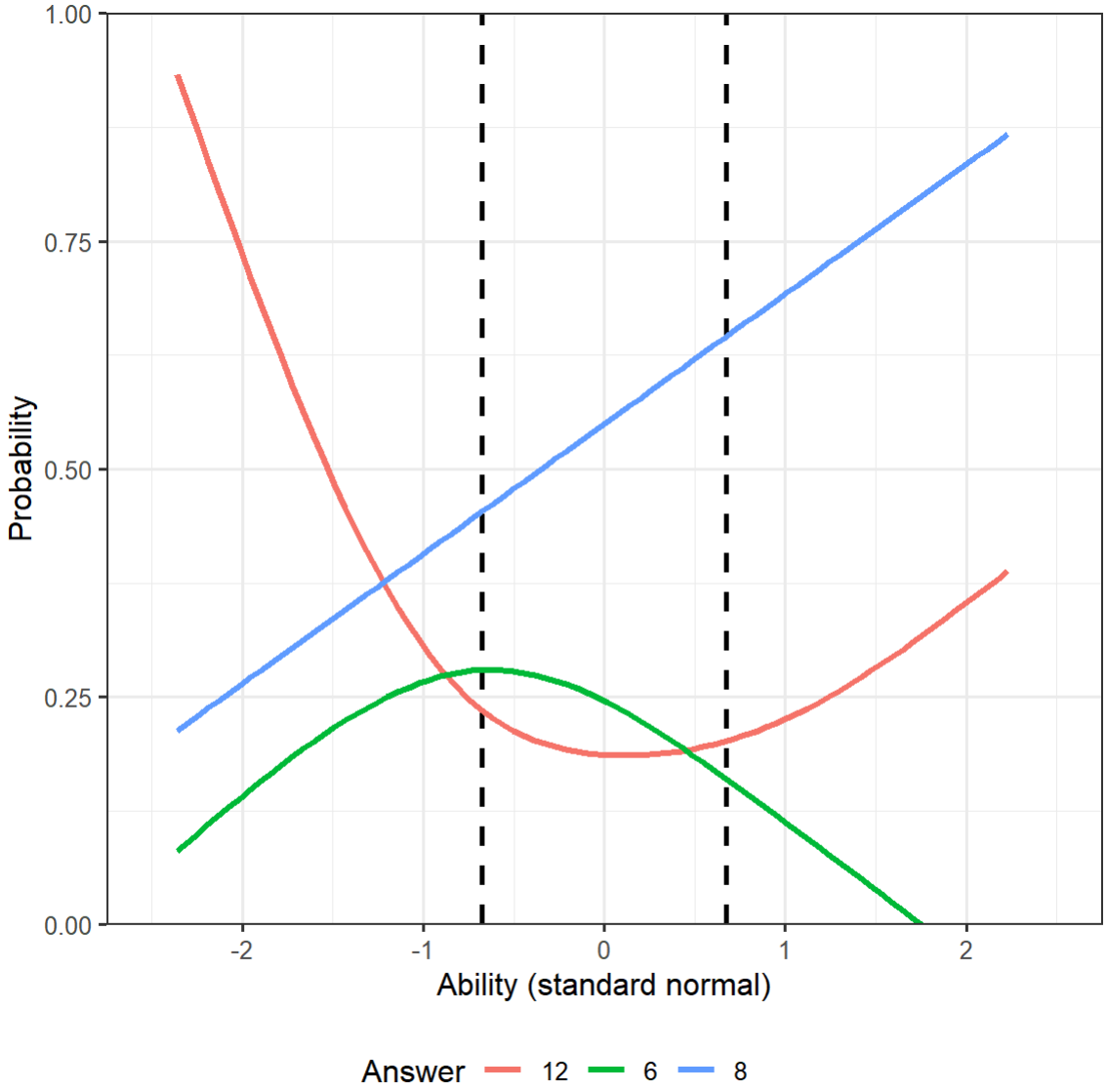
Answer	Frequency	Percent
6	41	13.5%
8	104	34.2%
10 (correct)	109	35.9%
12	50	16.4%
Overall	304	100.0%

10.1.26.4 Distractors functioning:

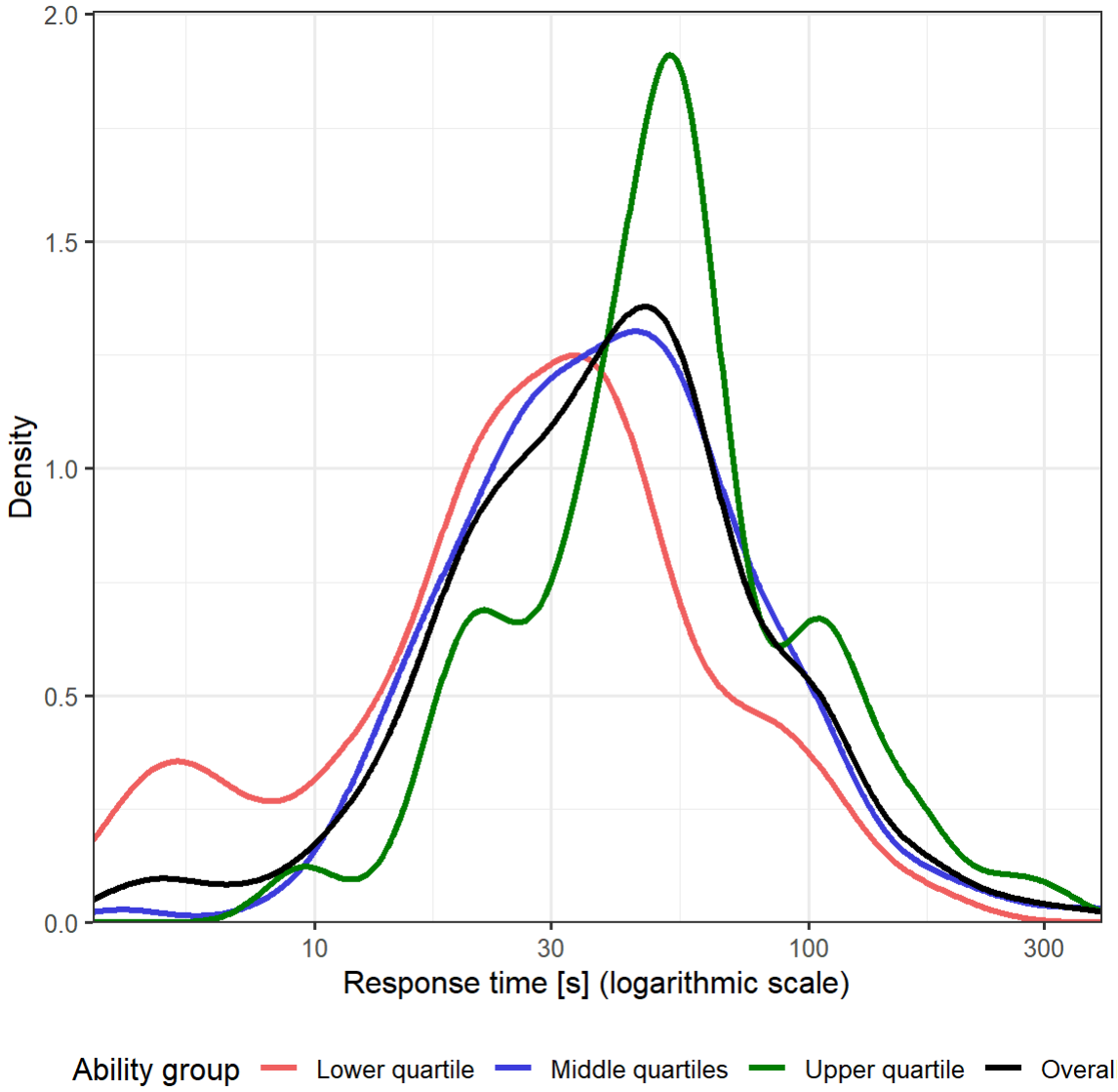
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
12	20	24	6	50	40.0%	22.0%	16.7%	25.6%
6	12	24	5	41	24.0%	22.0%	13.9%	21.0%
8	18	61	25	104	36.0%	56.0%	69.4%	53.3%
Overall	50	109	36	195	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



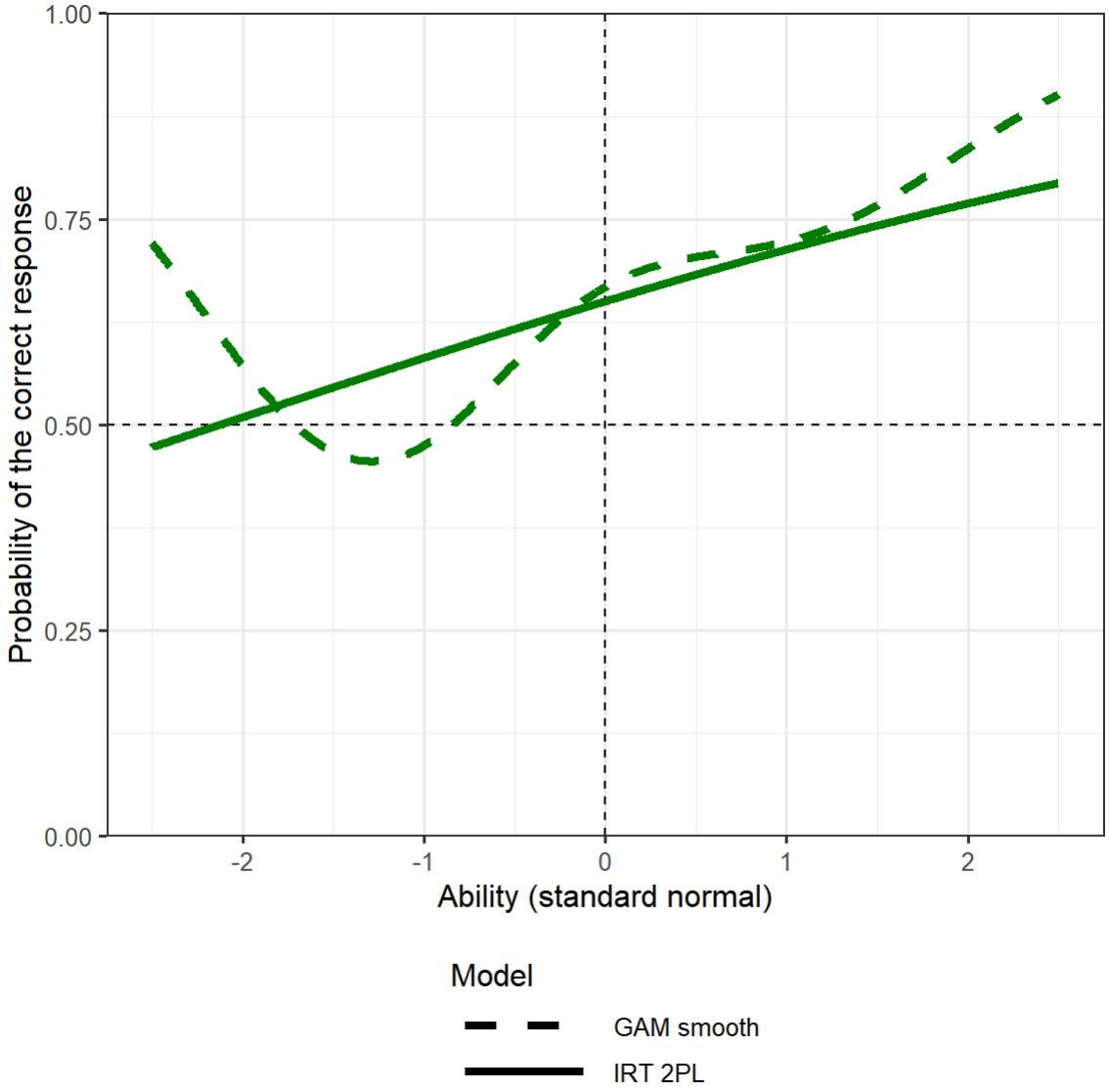
10.1.26.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	18.6	29.8	42.8	175.8	3.3	0.9
Middle quartiles	4.1	25.2	39.7	60.9	391.2	3.7	0.7
Upper quartile	9.2	34.6	52.4	65.8	306.0	3.9	0.7
Overall	3.5	24.0	40.3	59.9	391.2	3.6	0.8

10.1.27 M061274 (TIMSS)

10.1.27.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	64.47%	304	TIMSS	MC	0.18	0.18	0.29	−2.13	0.000

10.1.27.2 Gender DIF:

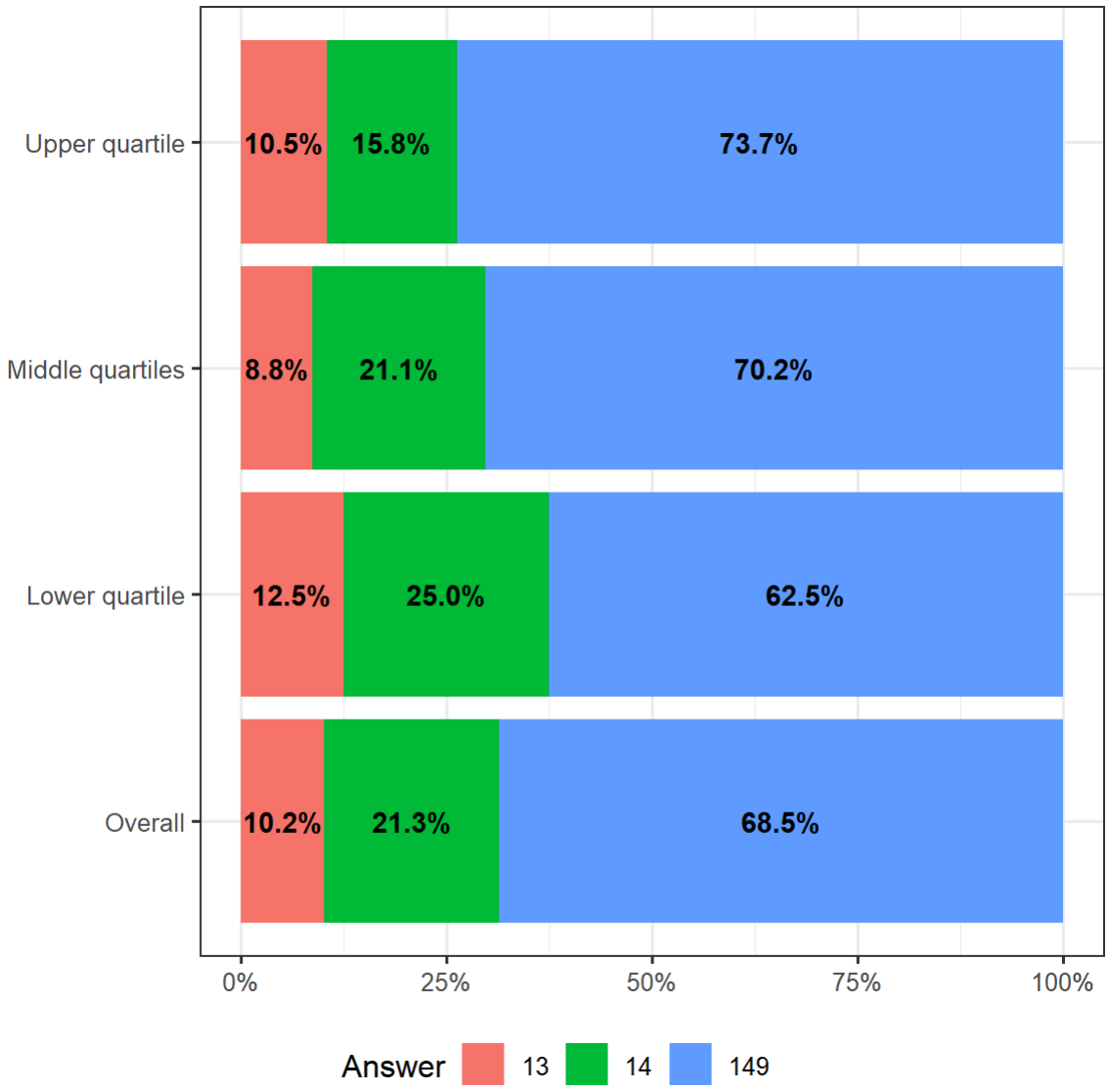
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.49	0.475	1.000	0.004	N/A

10.1.27.3 Frequency of answers:

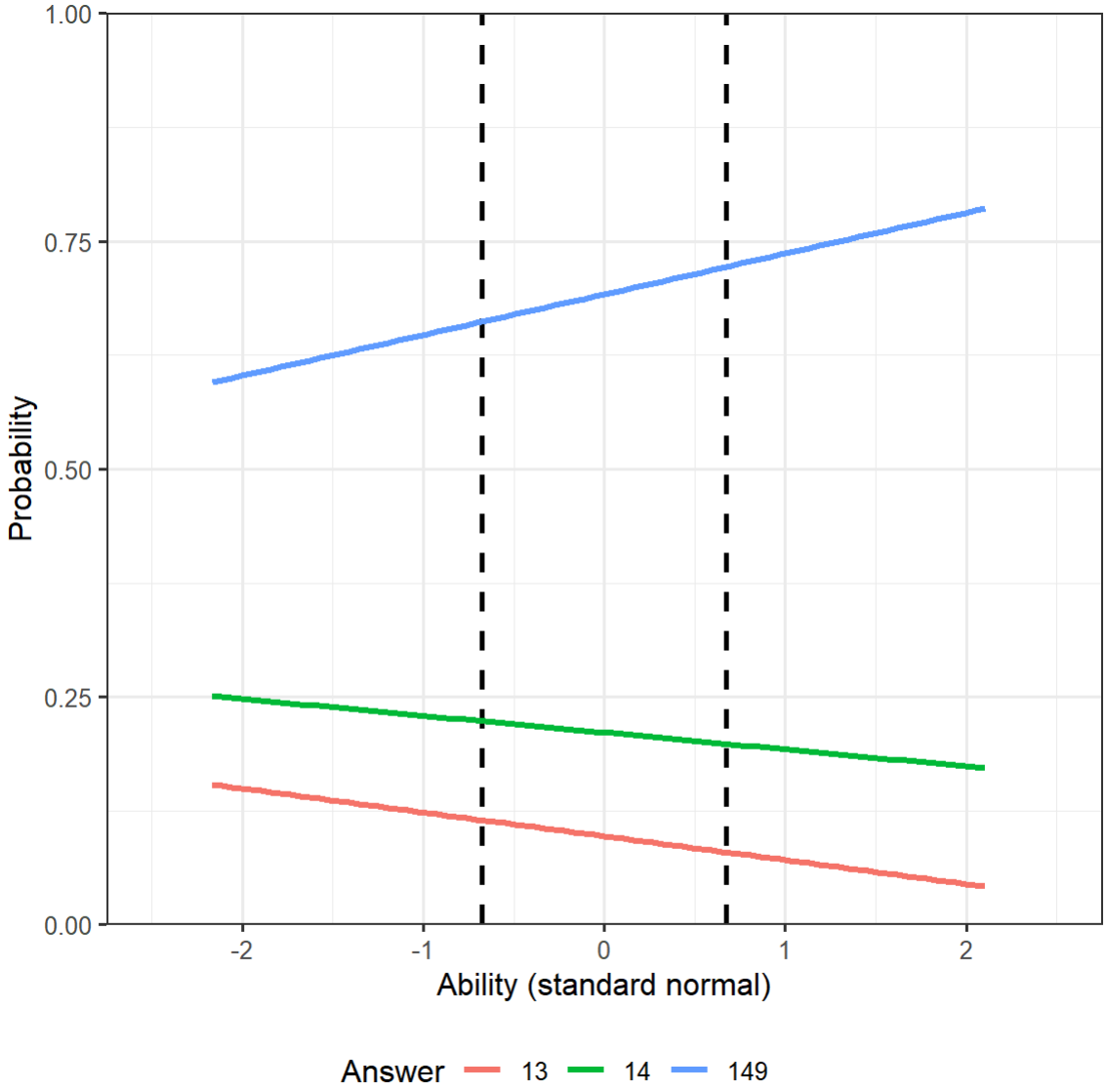
Answer	Frequency	Percent
13	11	3.6%
14	23	7.6%
134 (correct)	196	64.5%
149	74	24.3%
Overall	304	100.0%

10.1.27.4 Distractors functioning:

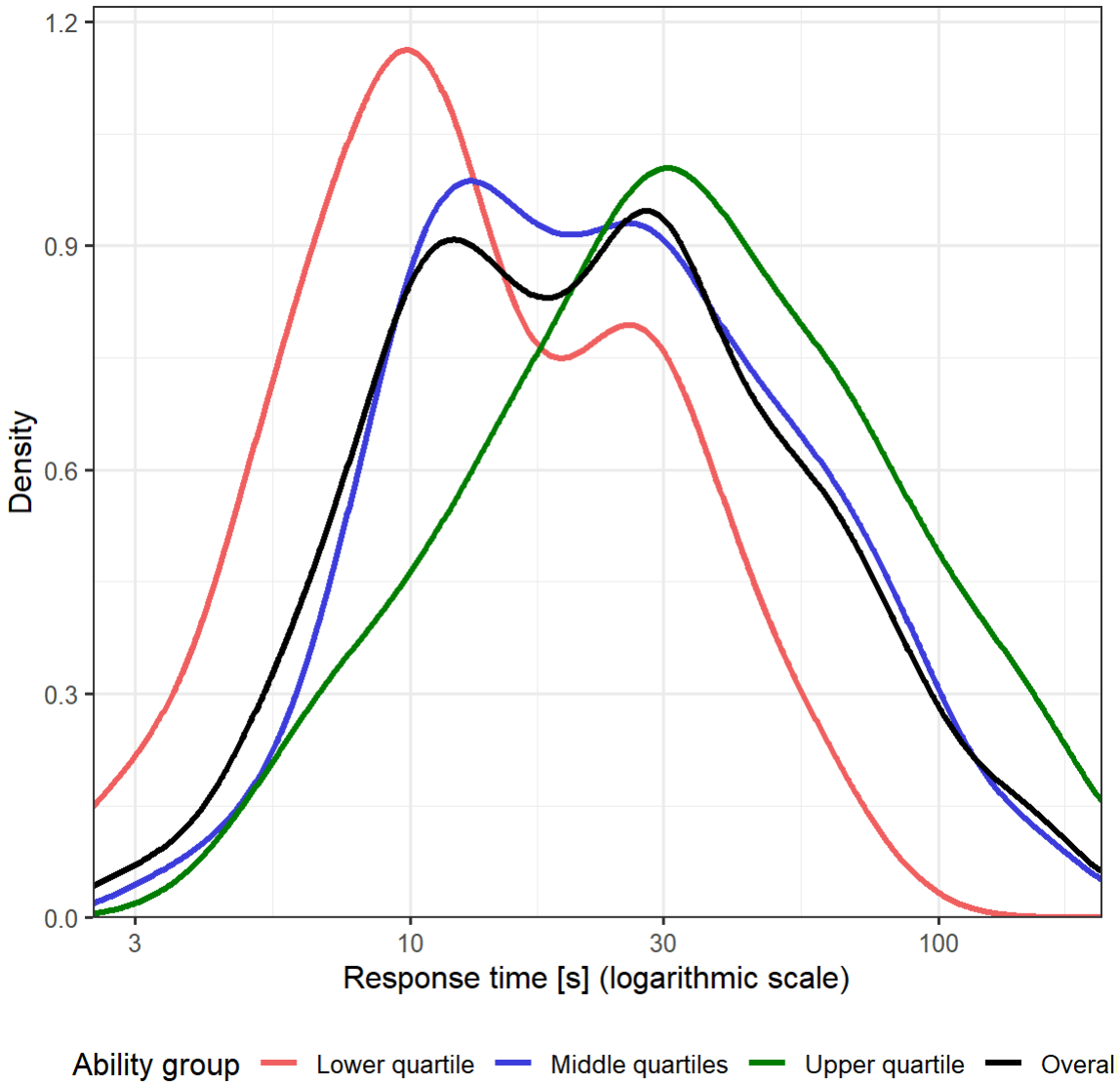
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
13	4	5	2	11	12.5%	8.8%	10.5%	10.2%
14	8	12	3	23	25.0%	21.1%	15.8%	21.3%
149	20	40	14	74	62.5%	70.2%	73.7%	68.5%
Overall	32	57	19	108	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



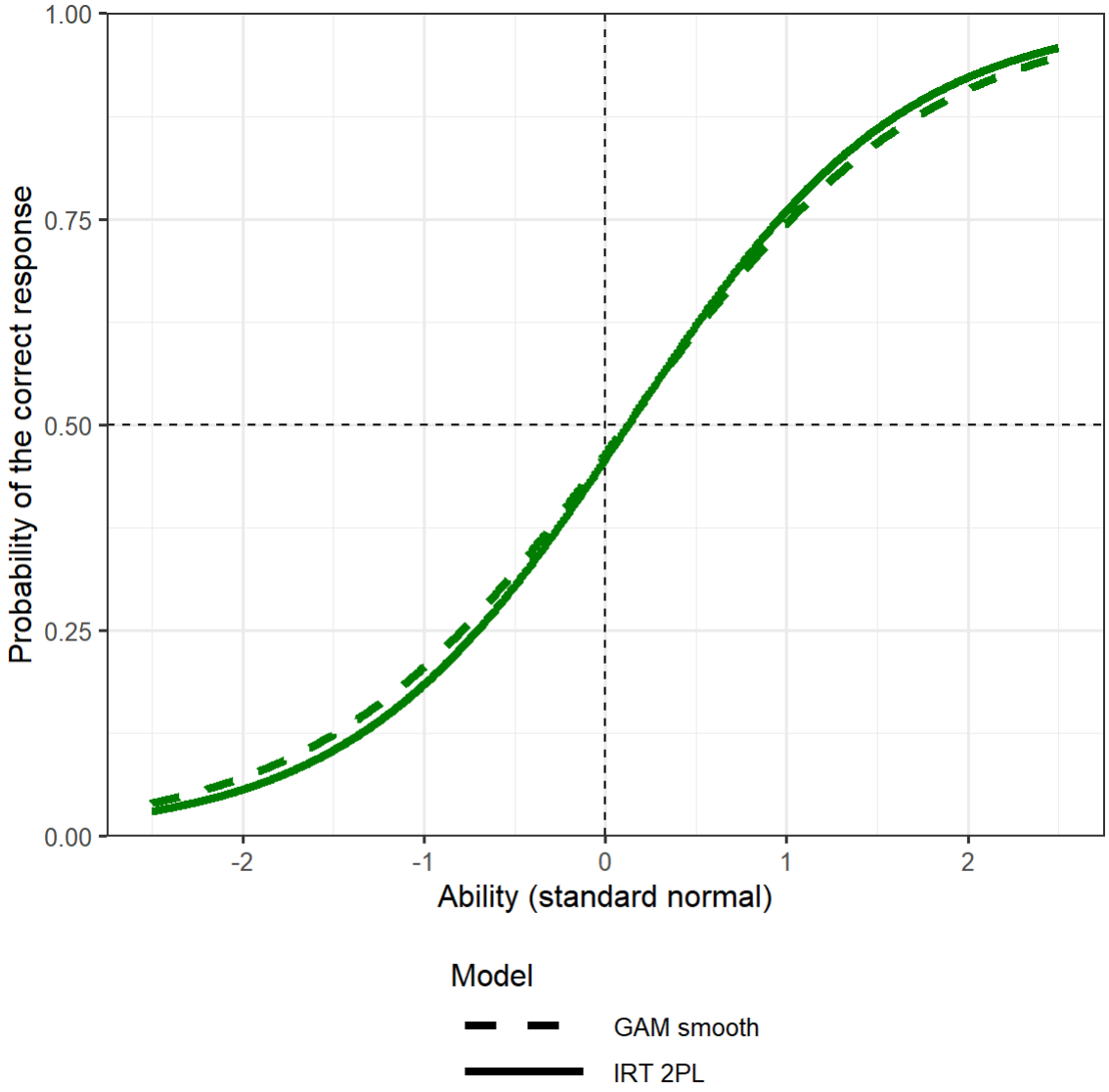
10.1.27.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.5	7.6	11.2	24.9	66.0	2.6	0.8
Middle quartiles	3.5	12.0	23.2	41.3	172.5	3.2	0.8
Upper quartile	5.0	16.2	30.7	58.5	202.8	3.5	0.9
Overall	2.5	11.2	22.9	40.2	202.8	3.1	0.9

10.1.28 M061207 (TIMSS)

10.1.28.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	47.99%	299	TIMSS	MC	0.47	0.50	1.32	0.13	0.015

10.1.28.2 Gender DIF:

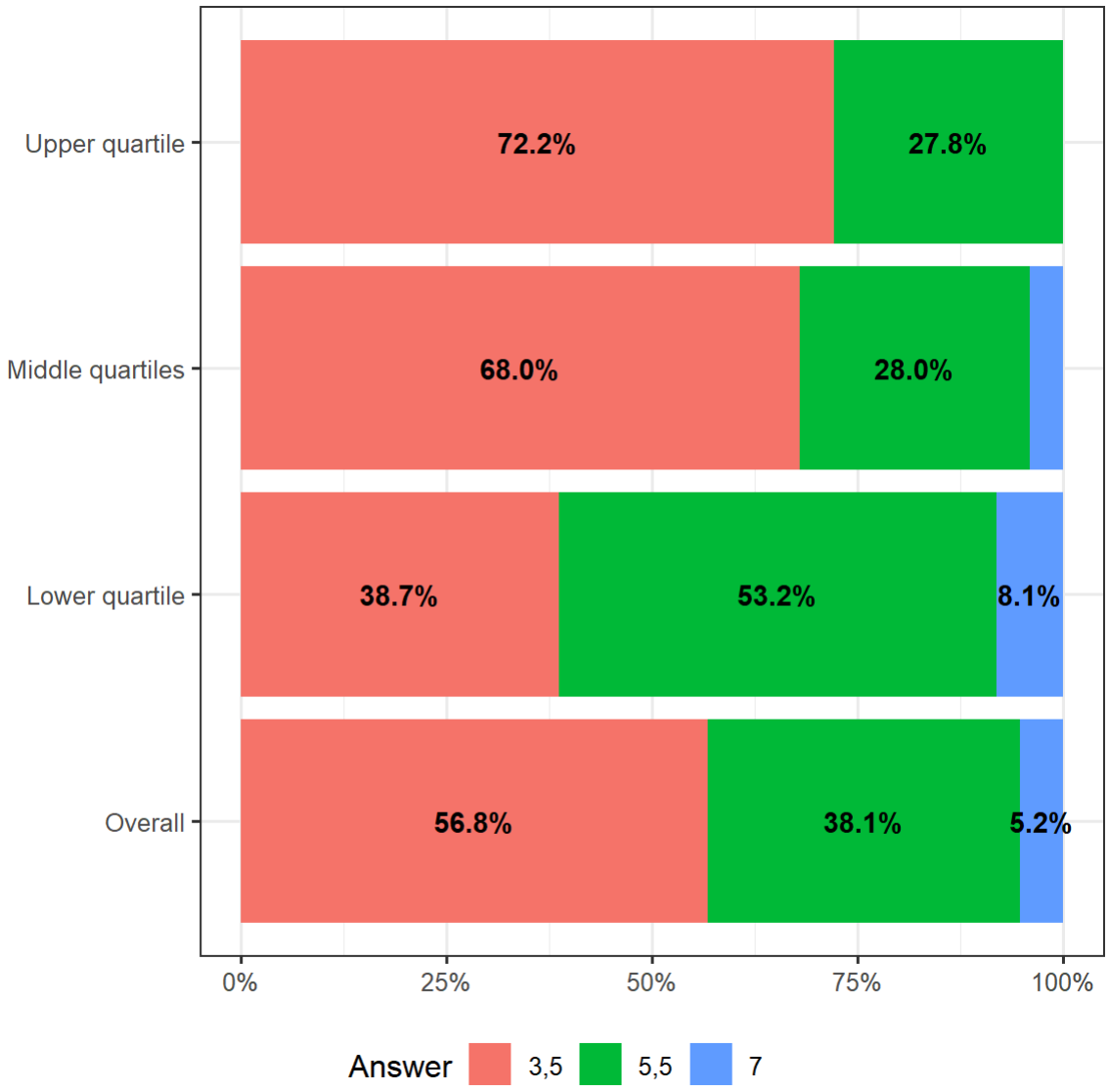
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.73	0.155	1.000	0.010	N/A

10.1.28.3 Frequency of answers:

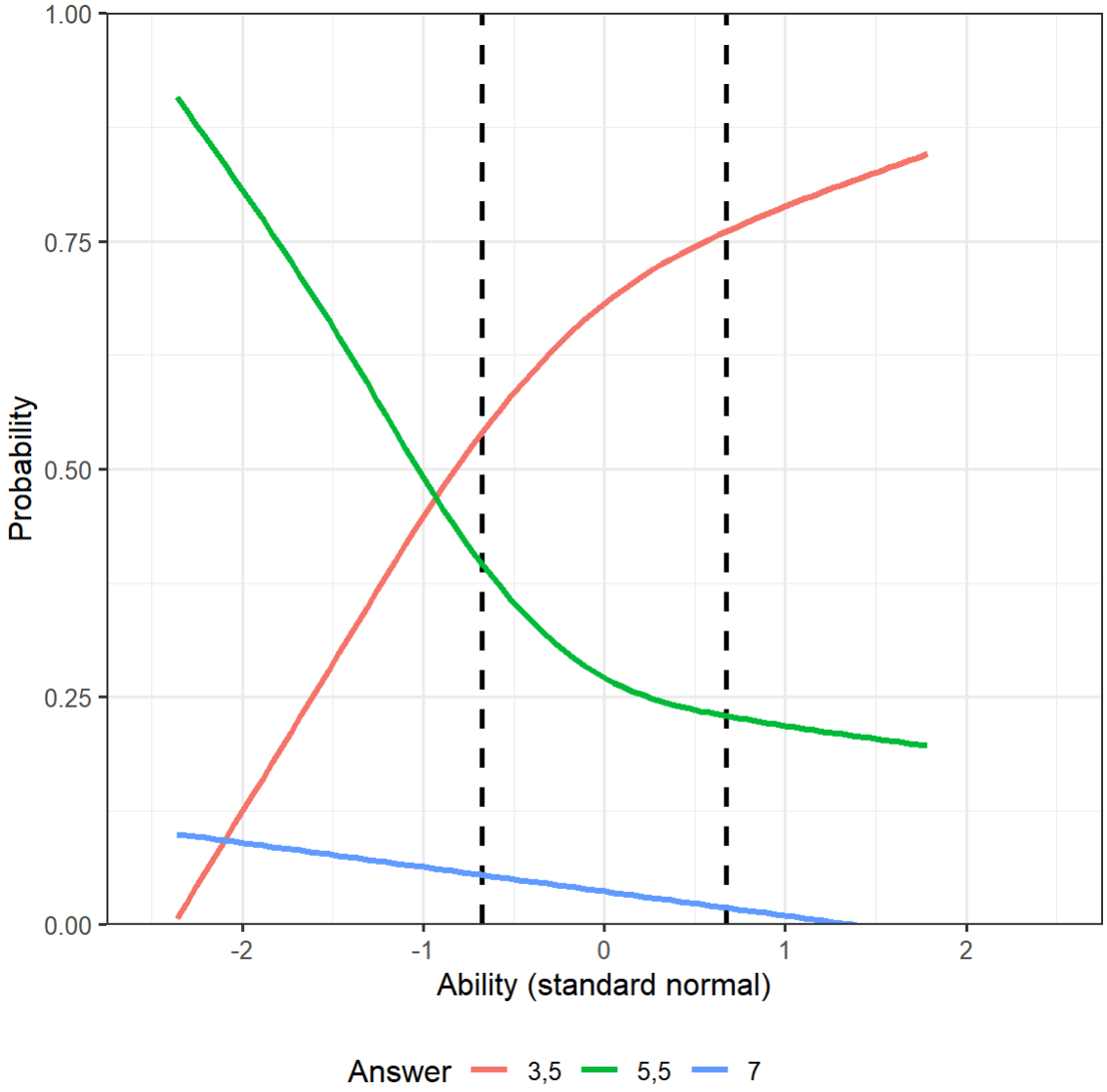
Answer	Frequency	Percent
3 (correct)	143	48.0%
3,5	88	29.5%
5,5	59	19.8%
7	8	2.7%
Overall	298	100.0%

10.1.28.4 Distractors functioning:

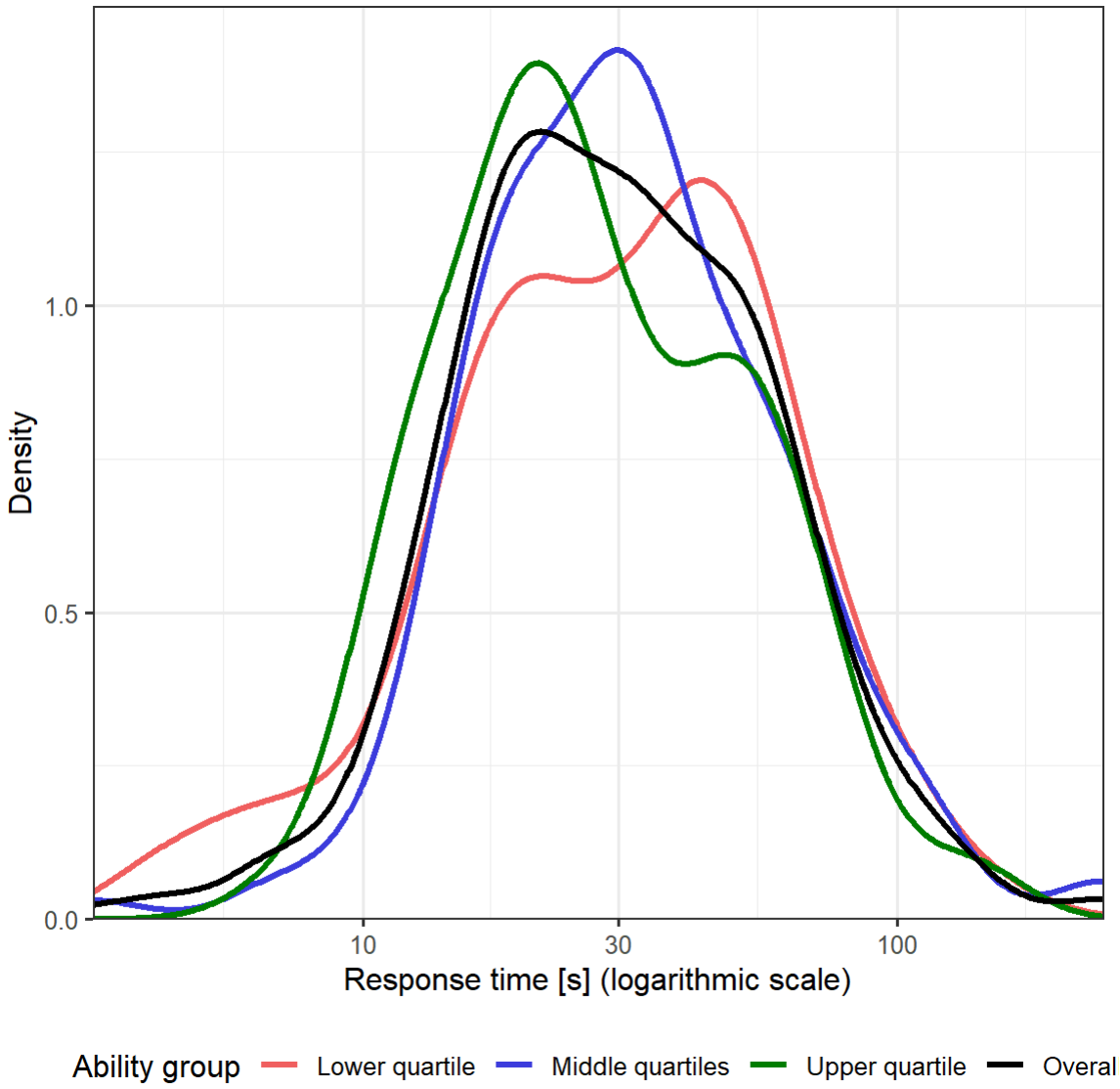
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
3,5	24	51	13	88	38.7%	68.0%	72.2%	56.8%
5,5	33	21	5	59	53.2%	28.0%	27.8%	38.1%
7	5	3	0	8	8.1%	4.0%	0.0%	5.2%
Overall	62	75	18	155	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



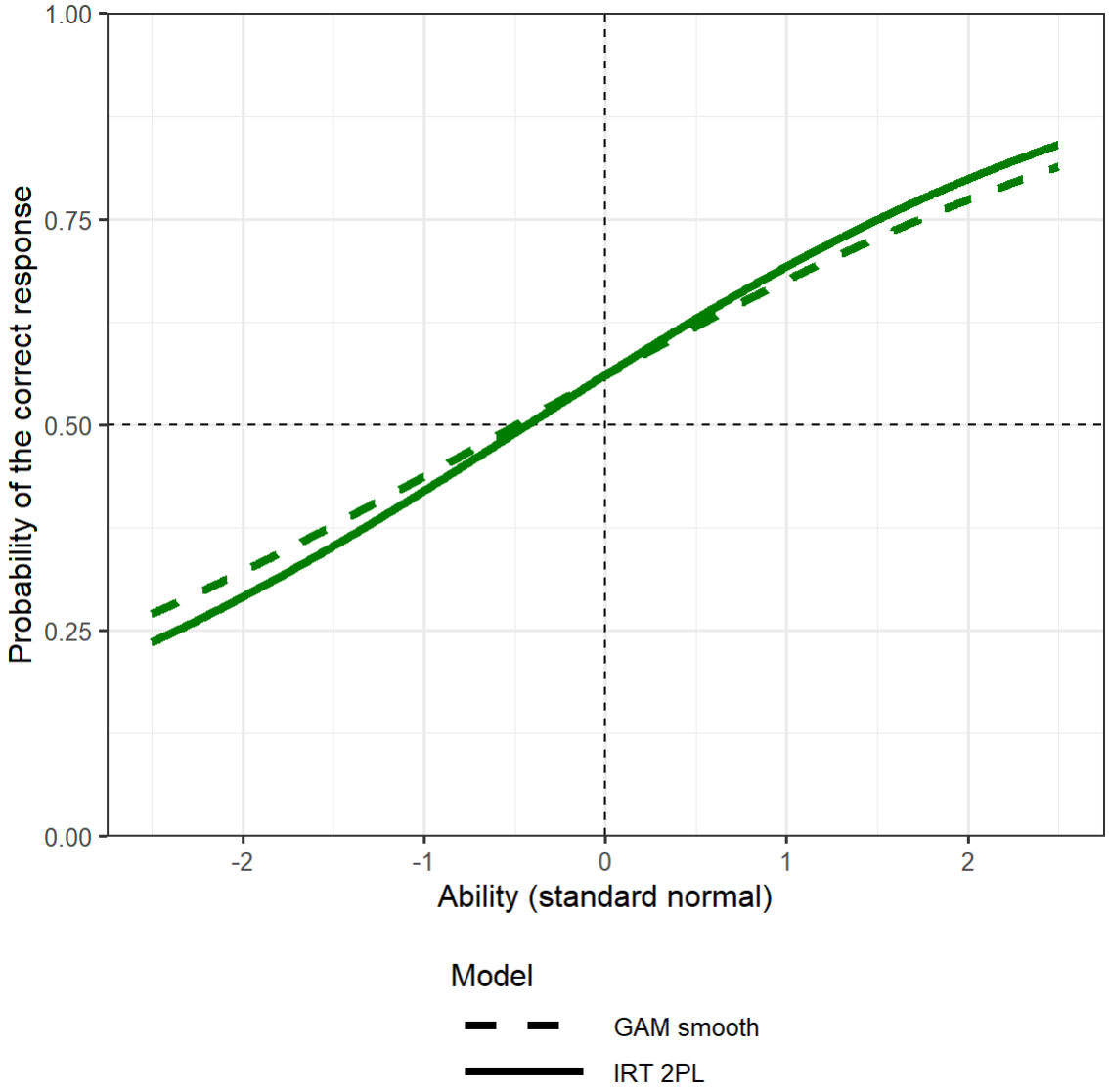
10.1.28.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.2	18.1	33.0	49.7	143.7	3.4	0.7
Middle quartiles	3.1	20.2	30.4	48.1	242.2	3.4	0.7
Upper quartile	7.0	17.8	24.9	45.5	137.8	3.3	0.6
Overall	3.1	18.5	29.1	48.8	242.2	3.4	0.7

10.1.29 M100010 (LLM)

10.1.29.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	55.37%	299	LLM simple setup	MC	0.24	0.15	0.57	−0.43	0.000

10.1.29.2 Gender DIF:

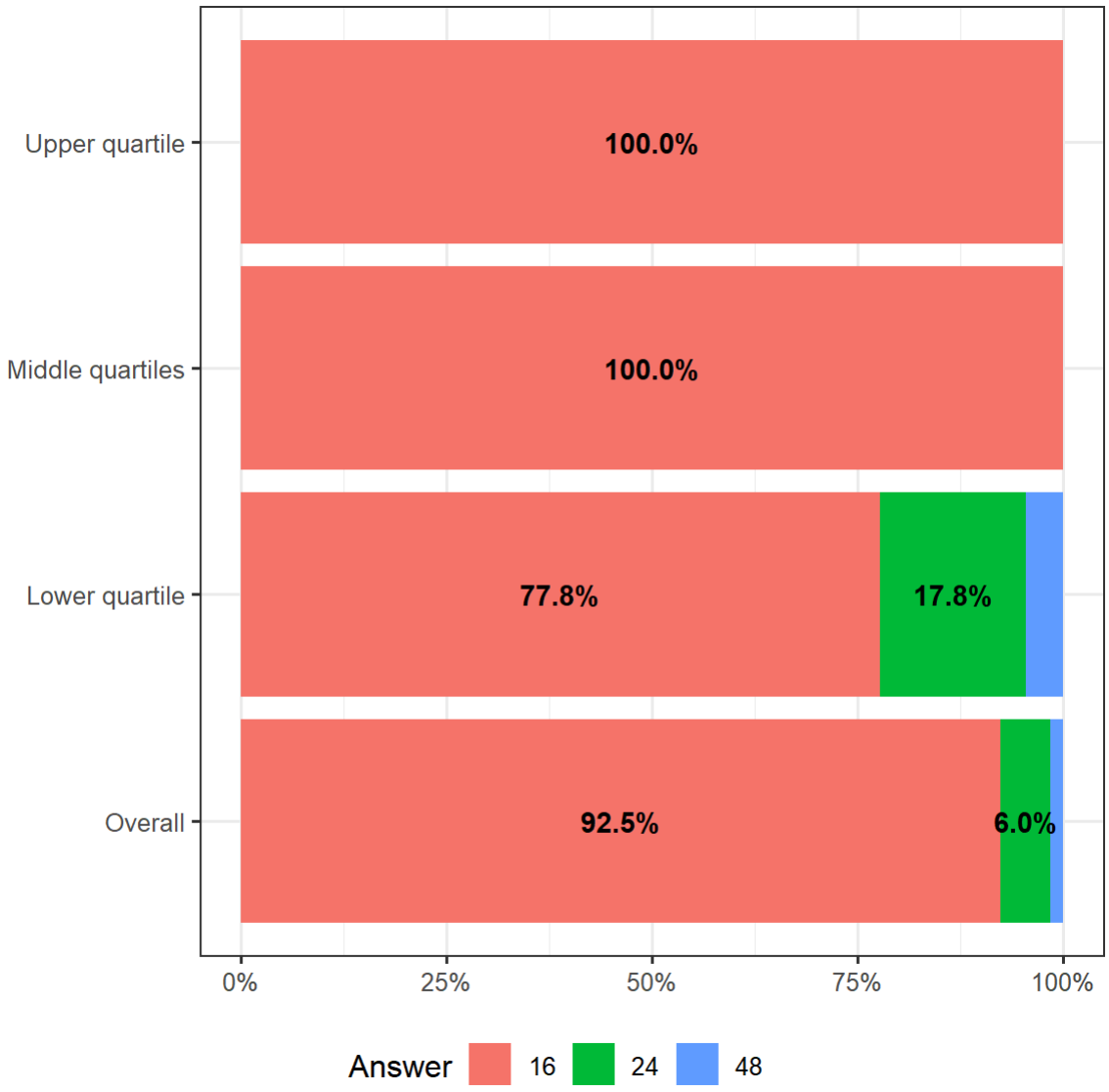
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
5.12	0.077	1.000	0.014	N/A

10.1.29.3 Frequency of answers:

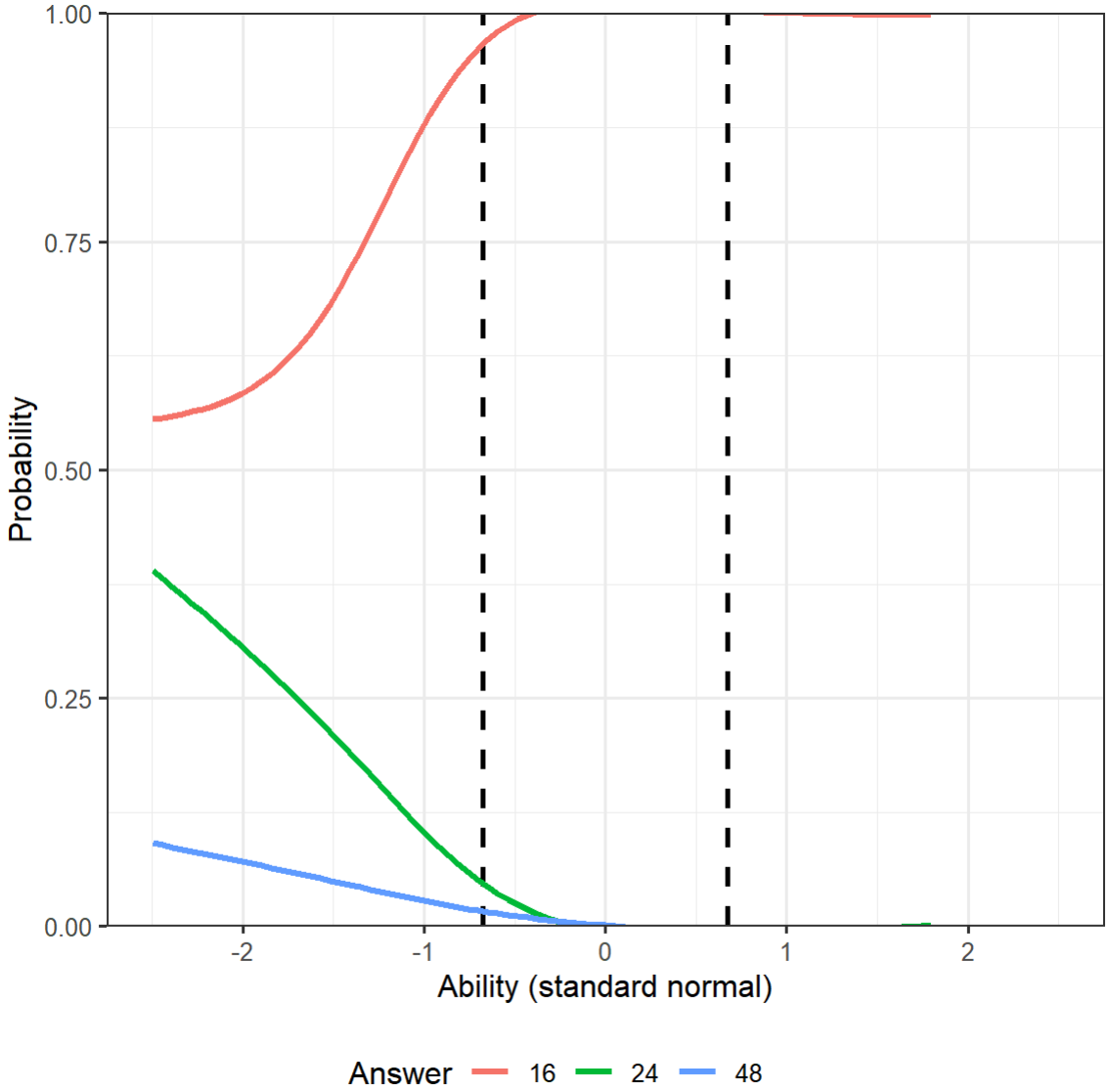
Answer	Frequency	Percent
12 (correct)	165	55.4%
16	123	41.3%
24	8	2.7%
48	2	0.7%
Overall	298	100.0%

10.1.29.4 Distractors functioning:

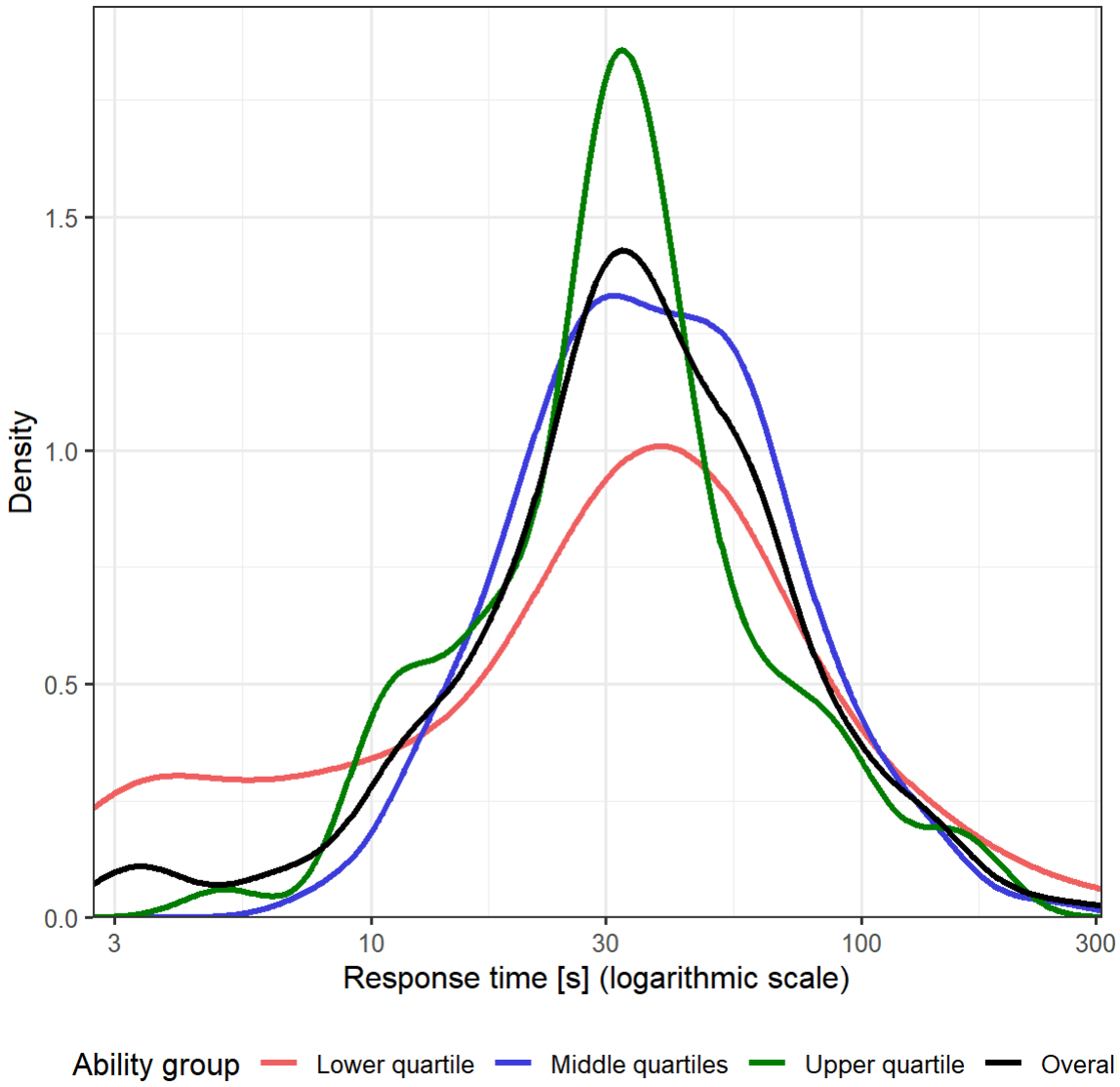
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
16	35	65	23	123	77.8%	100.0%	100.0%	92.5%
24	8	0	0	8	17.8%	0.0%	0.0%	6.0%
48	2	0	0	2	4.4%	0.0%	0.0%	1.5%
Overall	45	65	23	133	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



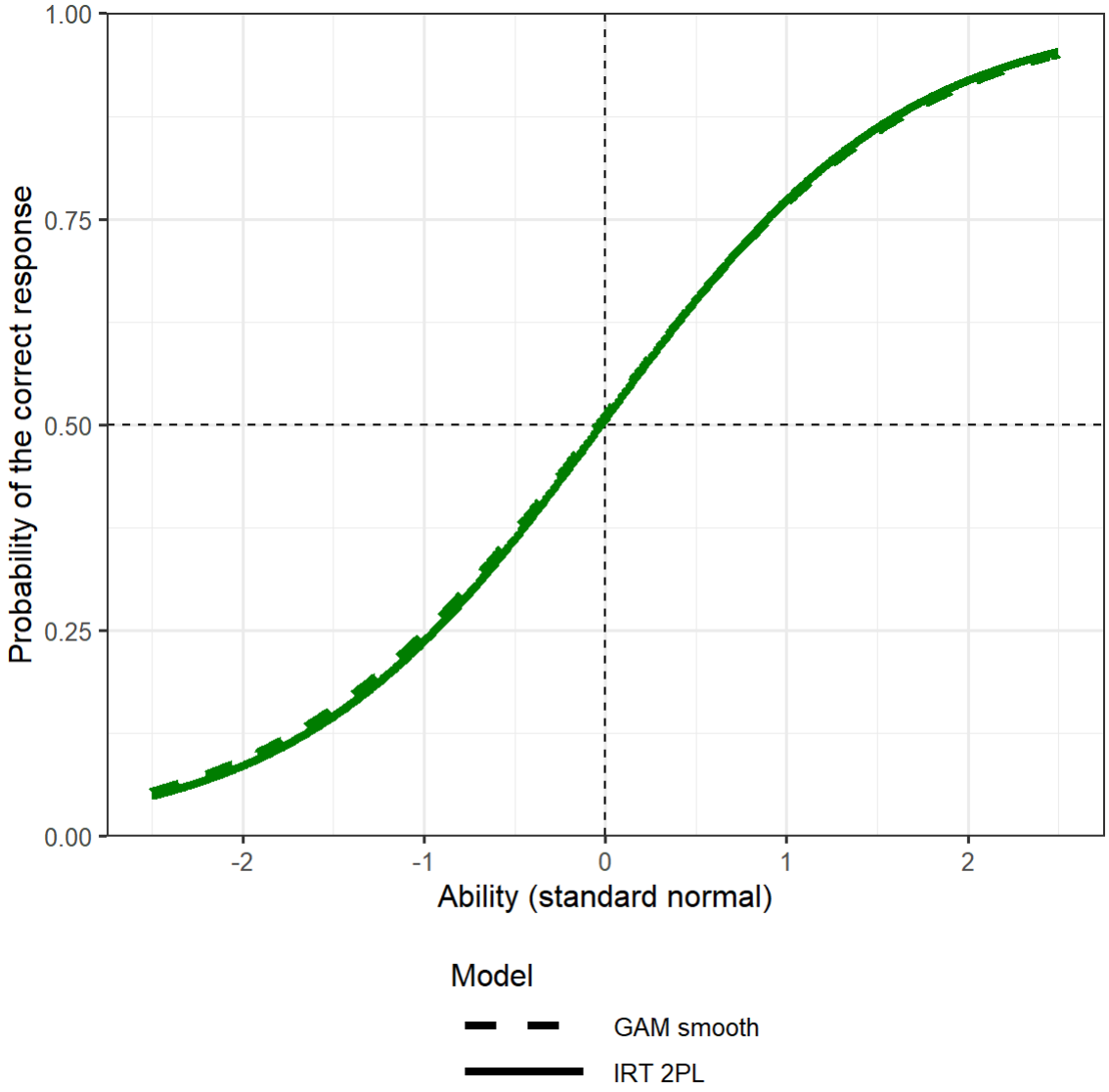
10.1.29.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	15.0	34.1	55.5	307.6	3.3	1.1
Middle quartiles	7.8	24.7	36.8	57.4	241.7	3.6	0.6
Upper quartile	5.0	22.2	31.7	44.6	173.9	3.5	0.7
Overall	2.7	22.9	33.8	55.5	307.6	3.5	0.8

10.1.30 MP71021 (TIMSS)

10.1.30.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	51.70%	294	TIMSS	MC	0.46	0.54	1.20	−0.02	0.000

10.1.30.2 Gender DIF:

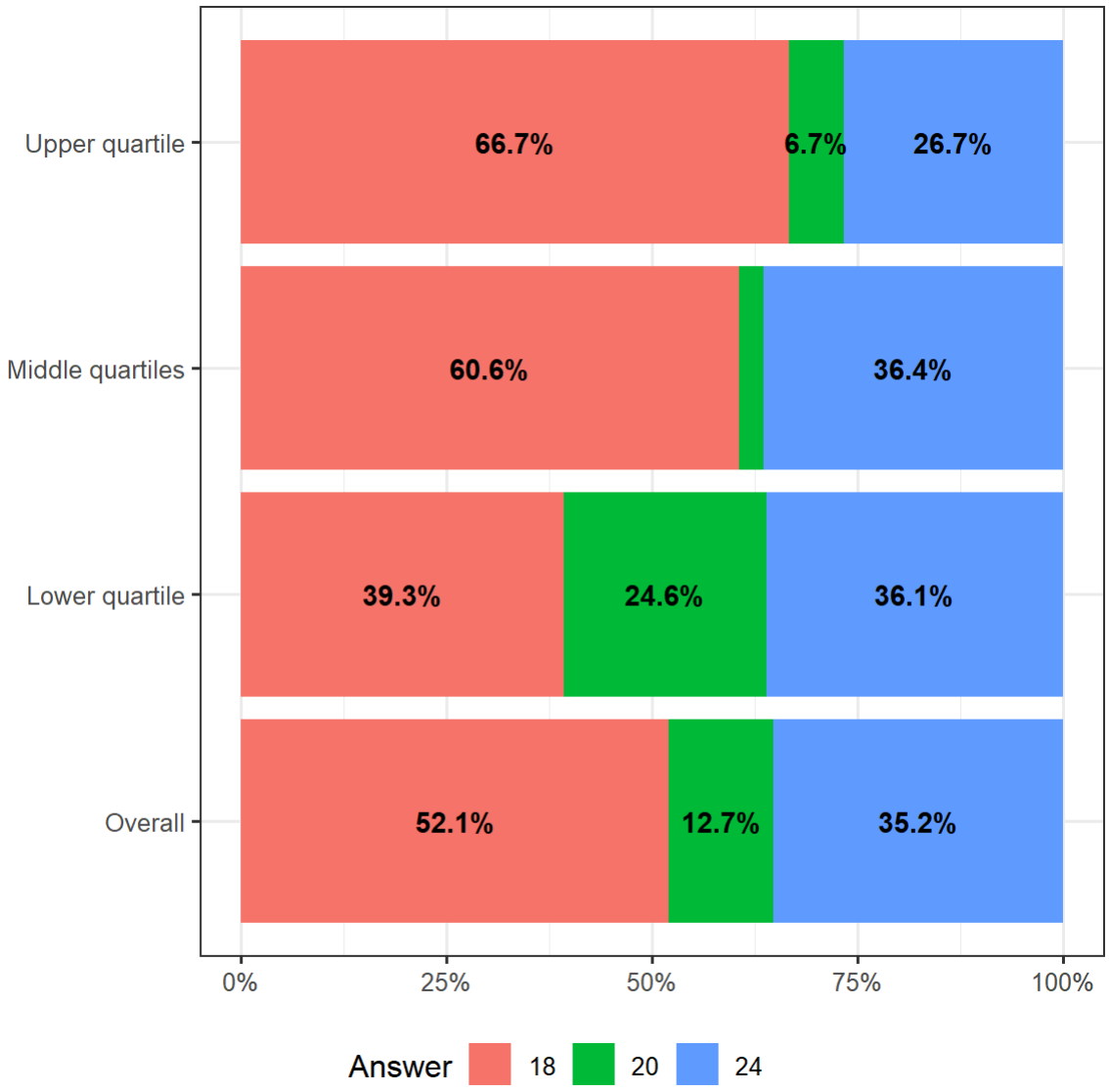
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.63	0.730	1.000	0.002	N/A

10.1.30.3 Frequency of answers:

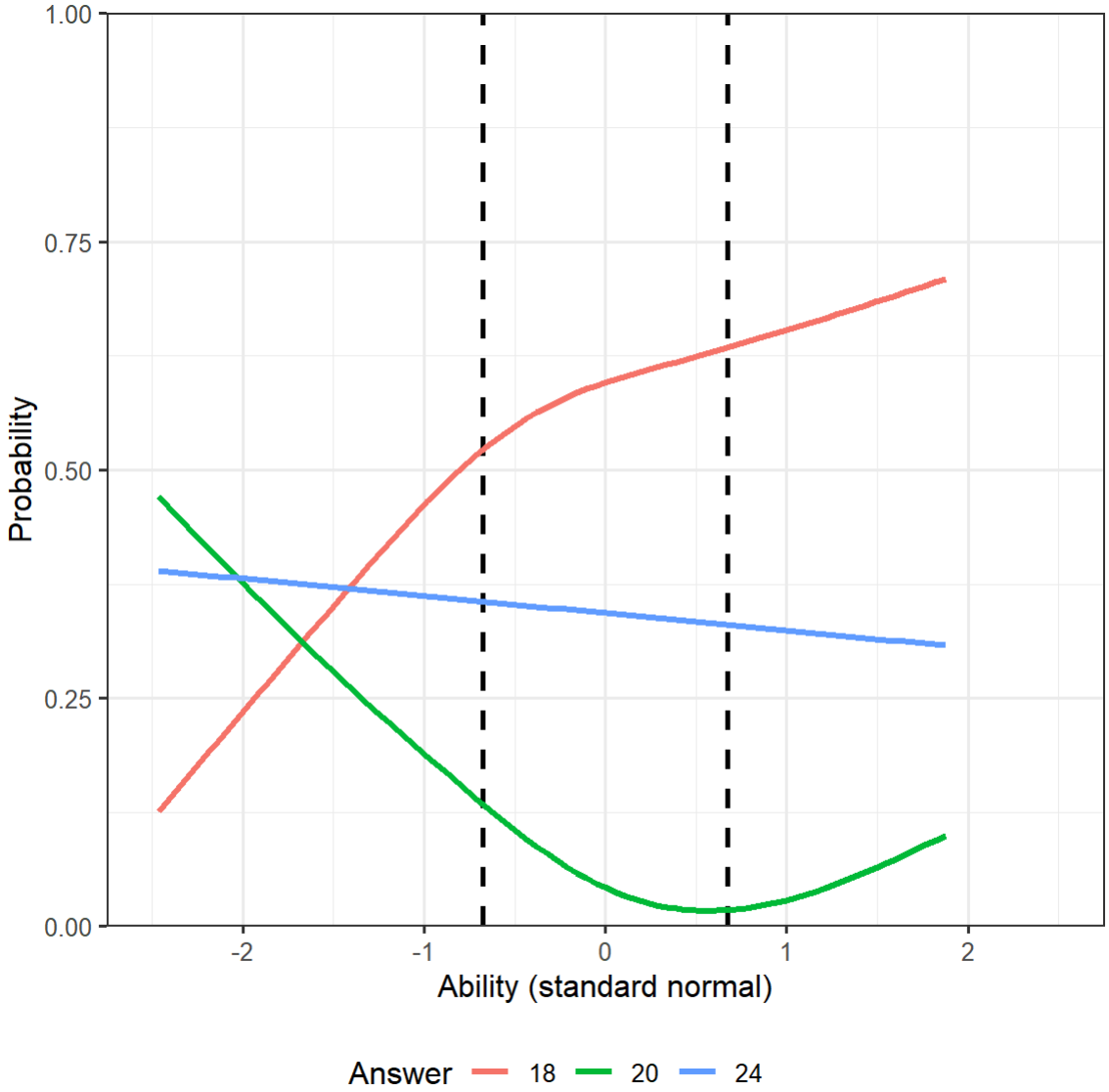
Answer	Frequency	Percent
18	74	25.2%
19 (correct)	152	51.7%
20	18	6.1%
24	50	17.0%
Overall	294	100.0%

10.1.30.4 Distractors functioning:

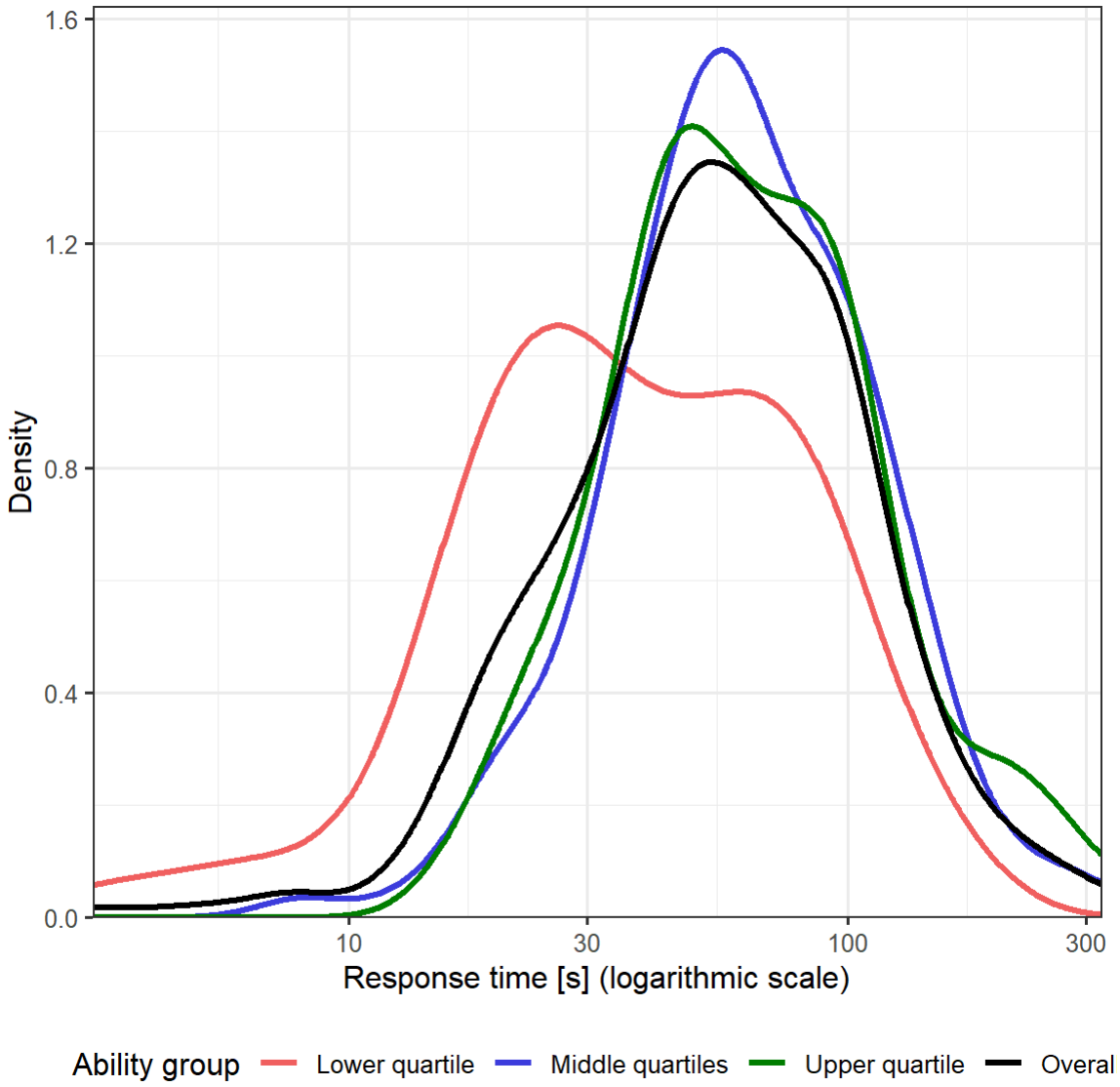
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
18	24	40	10	74	39.3%	60.6%	66.7%	52.1%
20	15	2	1	18	24.6%	3.0%	6.7%	12.7%
24	22	24	4	50	36.1%	36.4%	26.7%	35.2%
Overall	61	66	15	142	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



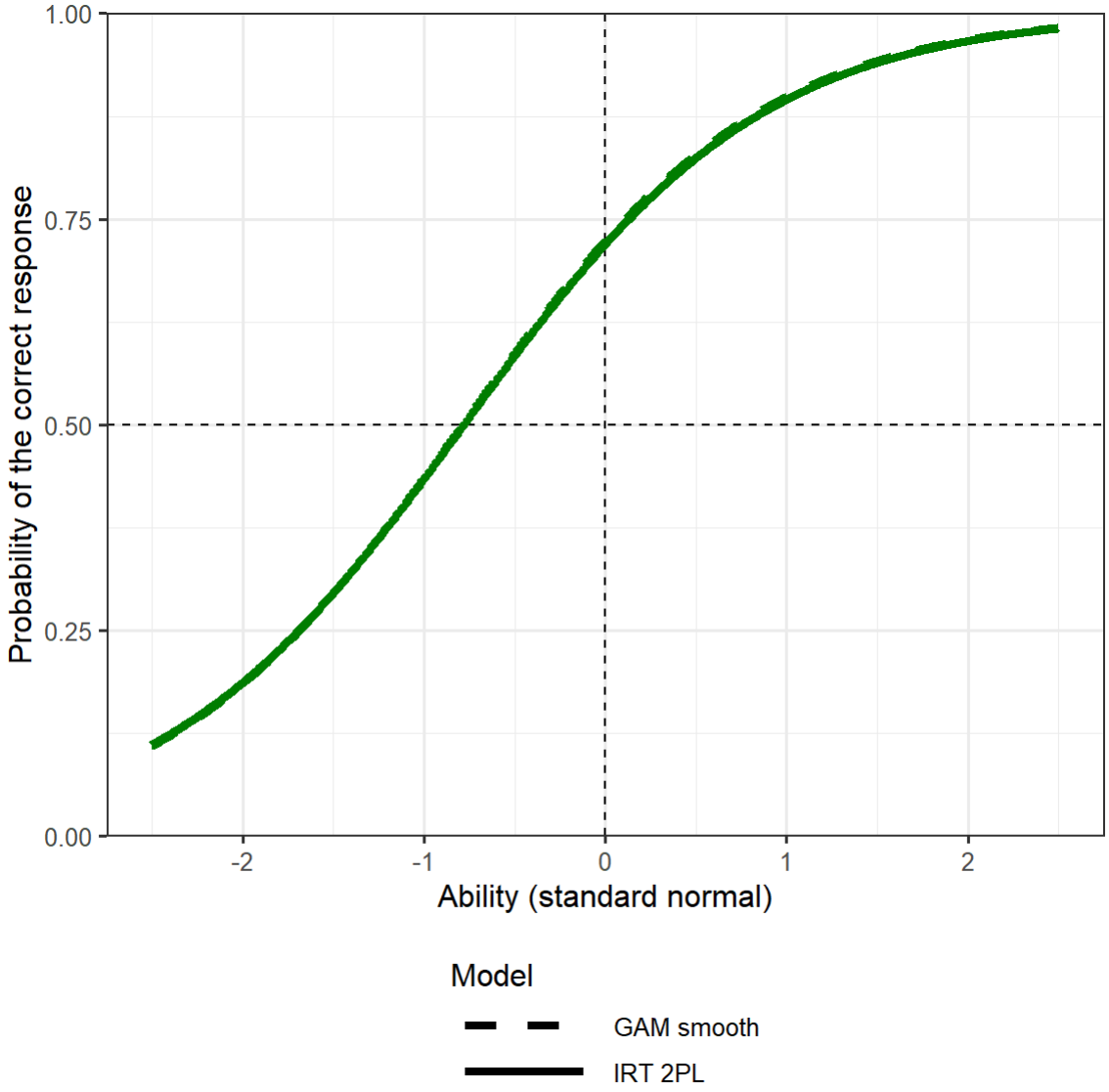
10.1.30.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	21.9	35.2	69.4	161.2	3.6	0.8
Middle quartiles	8.0	42.1	60.9	92.4	300.9	4.1	0.6
Upper quartile	17.0	40.8	62.1	93.9	321.9	4.1	0.6
Overall	3.1	36.0	55.3	88.4	321.9	4.0	0.7

10.1.31 MP71090 (TIMSS)

10.1.31.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	68.46%	299	TIMSS	MC	0.45	0.48	1.21	−0.78	0.011

10.1.31.2 Gender DIF:

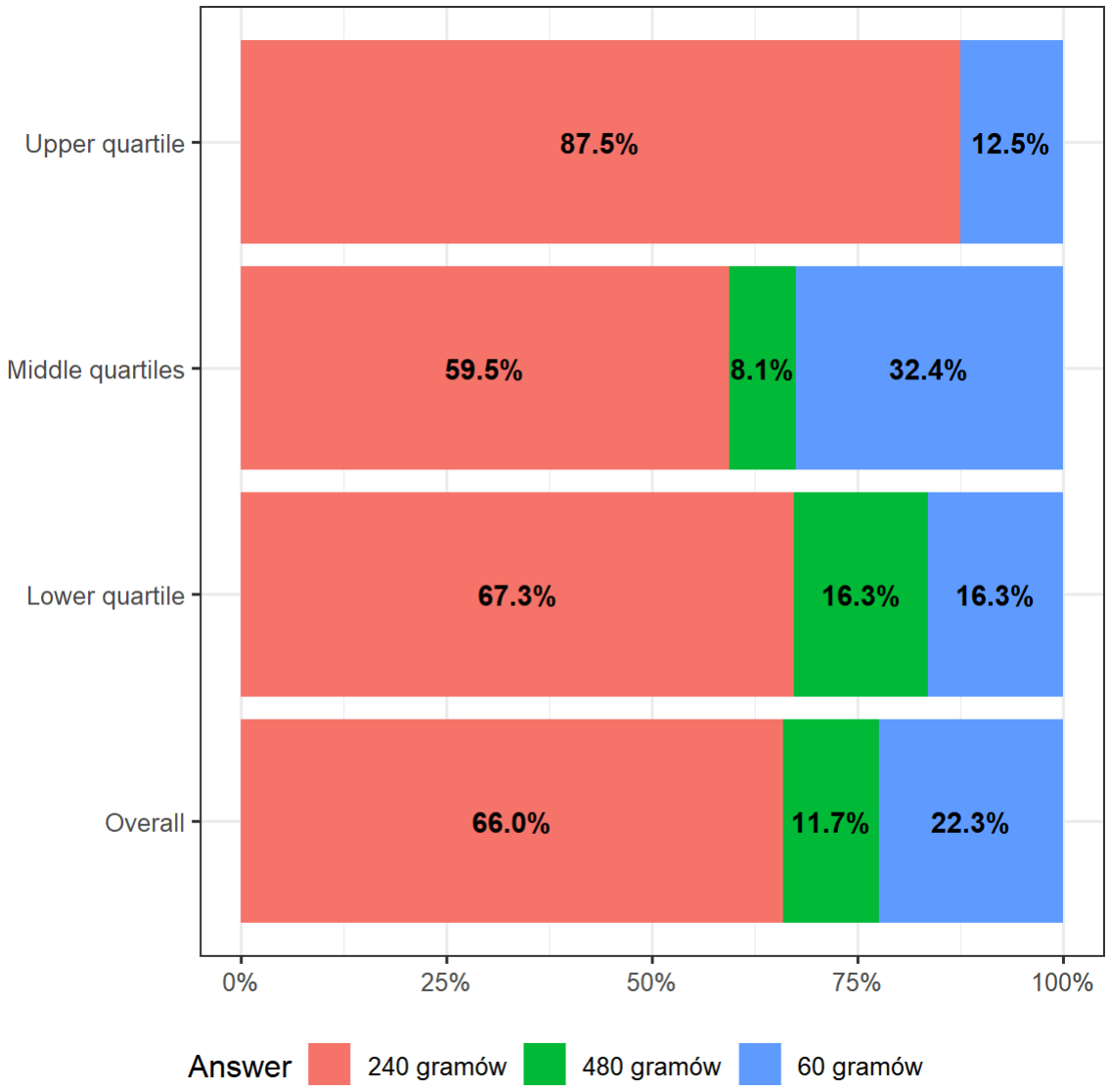
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.40	0.817	1.000	0.001	N/A

10.1.31.3 Frequency of answers:

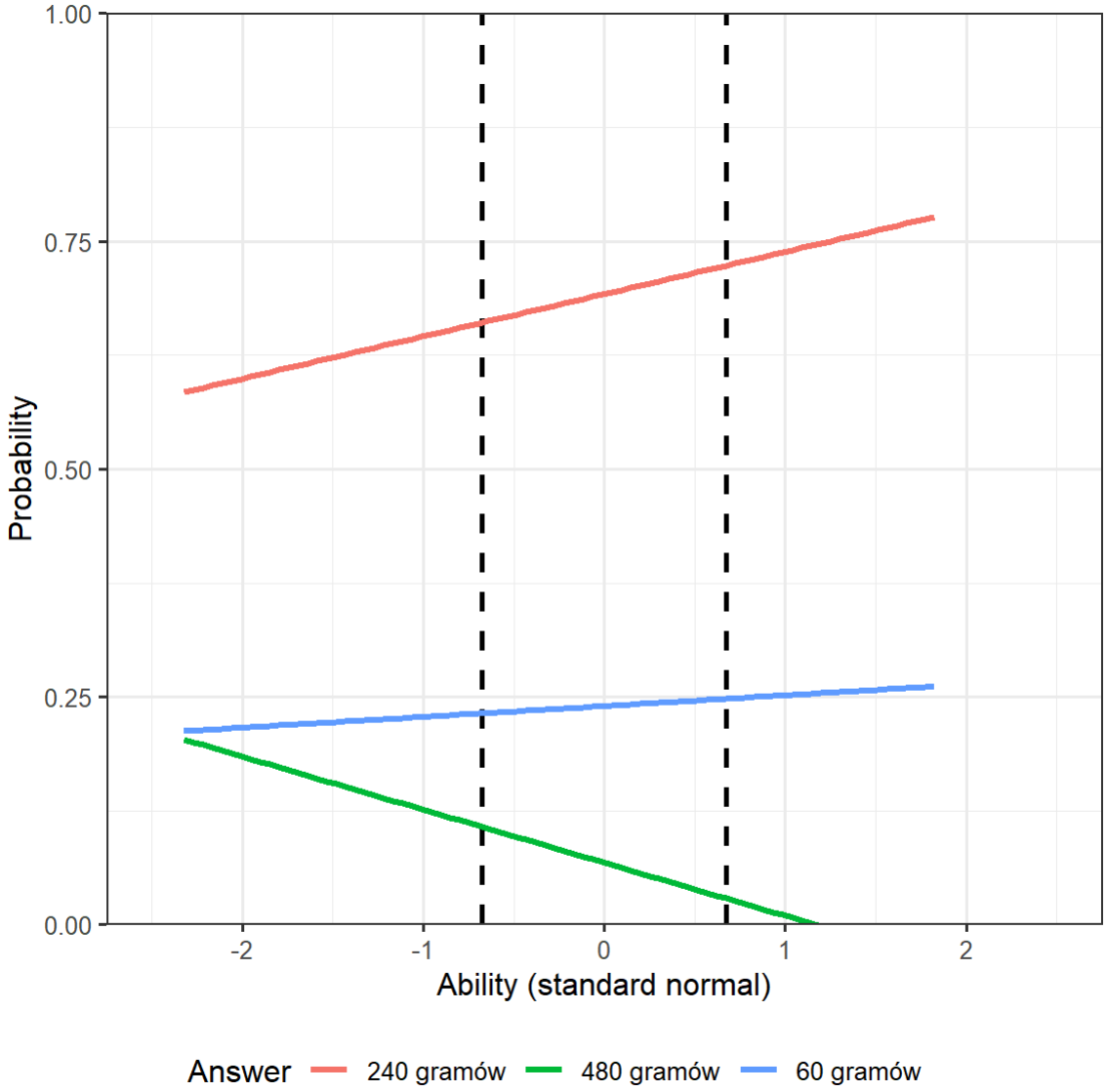
Answer	Frequency	Percent
120 gramów (correct)	204	68.5%
240 gramów	62	20.8%
480 gramów	11	3.7%
60 gramów	21	7.0%
Overall	298	100.0%

10.1.31.4 Distractors functioning:

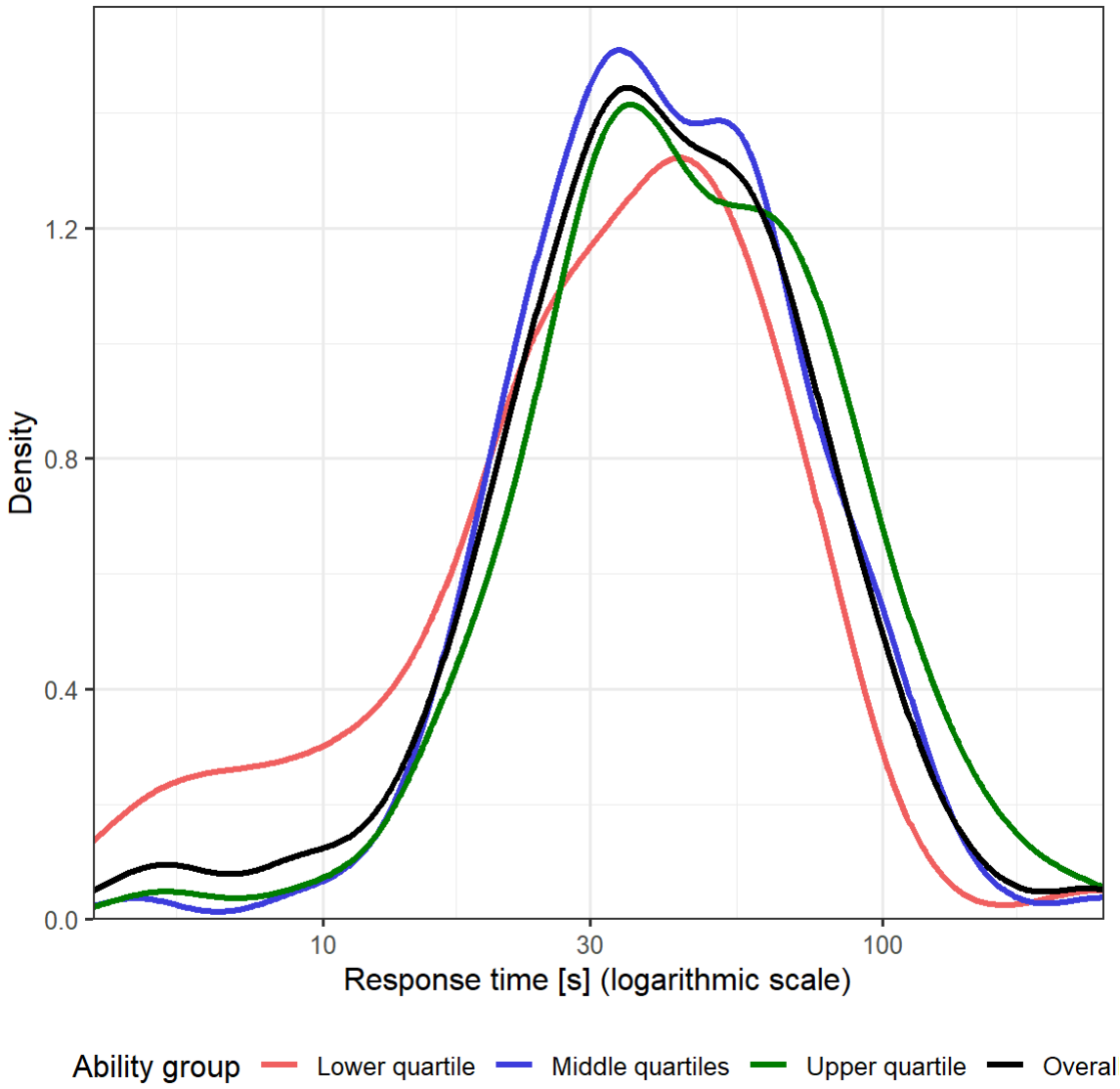
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
240 gramów	33	22	7	62	67.3%	59.5%	87.5%	66.0%
480 gramów	8	3	0	11	16.3%	8.1%	0.0%	11.7%
60 gramów	8	12	1	21	16.3%	32.4%	12.5%	22.3%
Overall	49	37	8	94	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



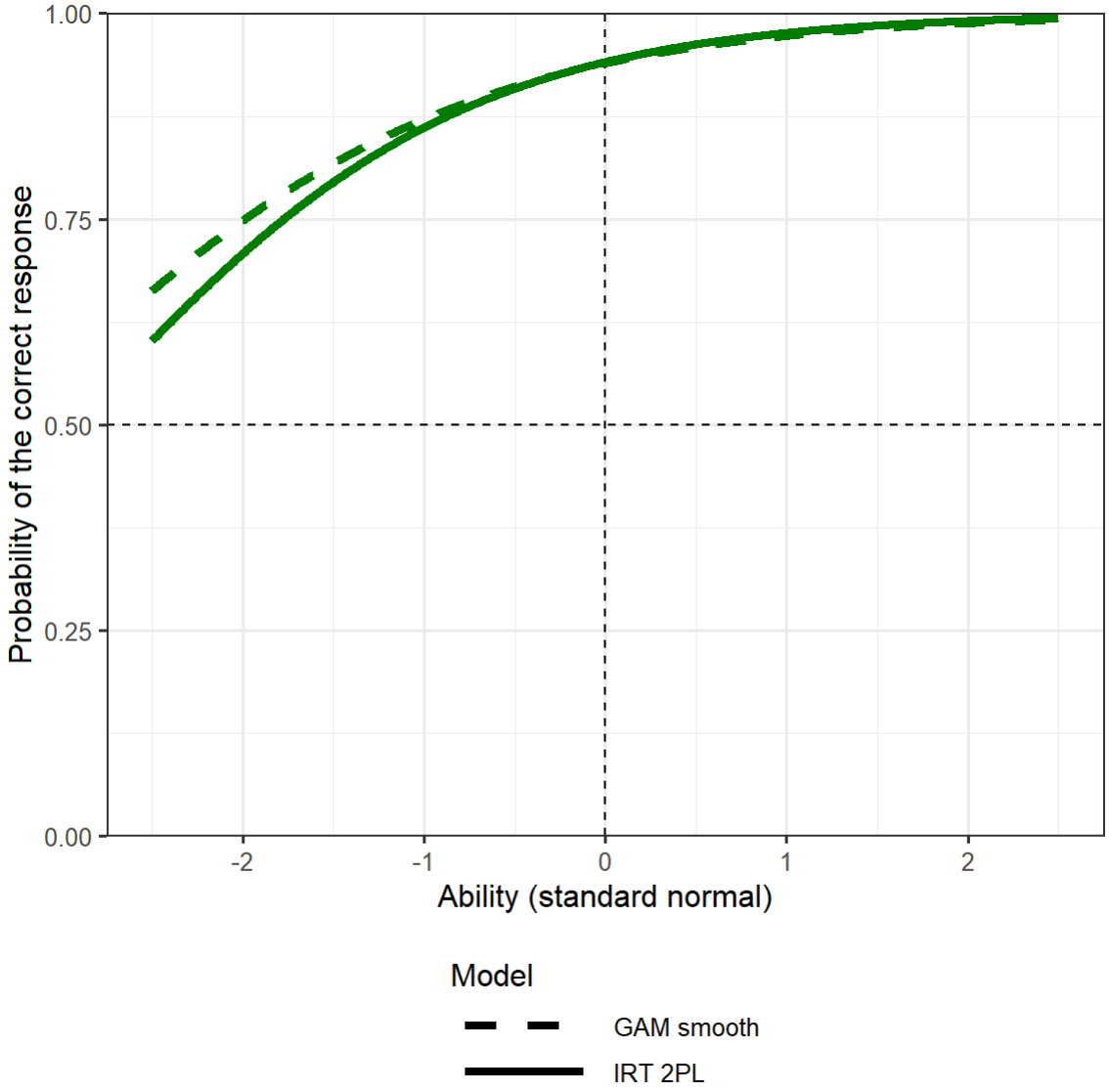
10.1.31.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.9	21.5	34.7	50.9	241.3	3.4	0.8
Middle quartiles	4.6	28.3	40.2	58.6	247.2	3.7	0.6
Upper quartile	5.2	30.0	43.3	68.7	227.4	3.8	0.6
Overall	3.9	26.3	38.8	60.2	247.2	3.7	0.7

10.1.32 M090012 (LLM)

10.1.32.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	91.79%	281	LLM teachers' selection	MC	0.24	0.13	0.94	−2.94	0.0(

10.1.32.2 Gender DIF:

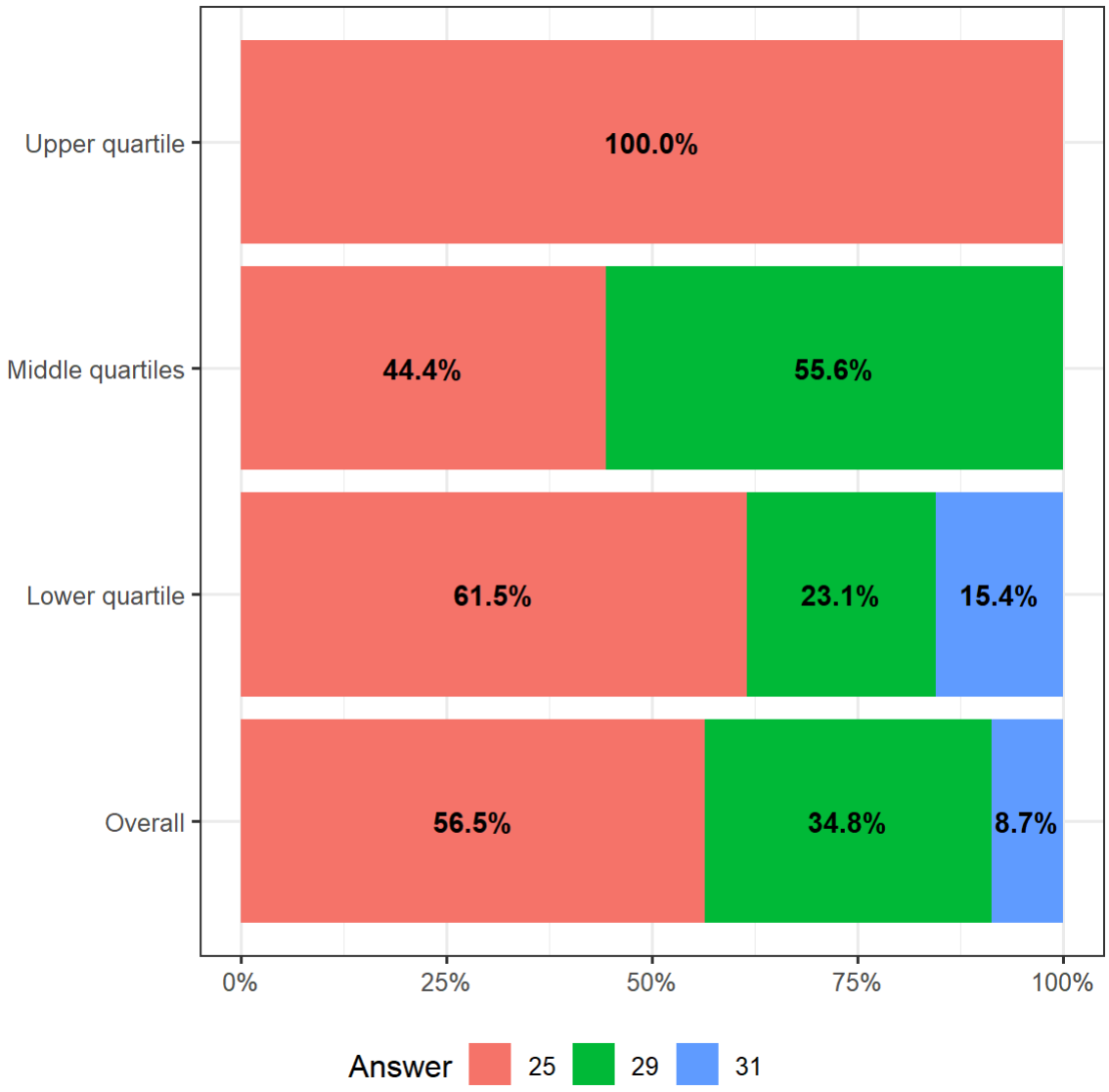
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.58	0.750	1.000	0.004	N/A

10.1.32.3 Frequency of answers:

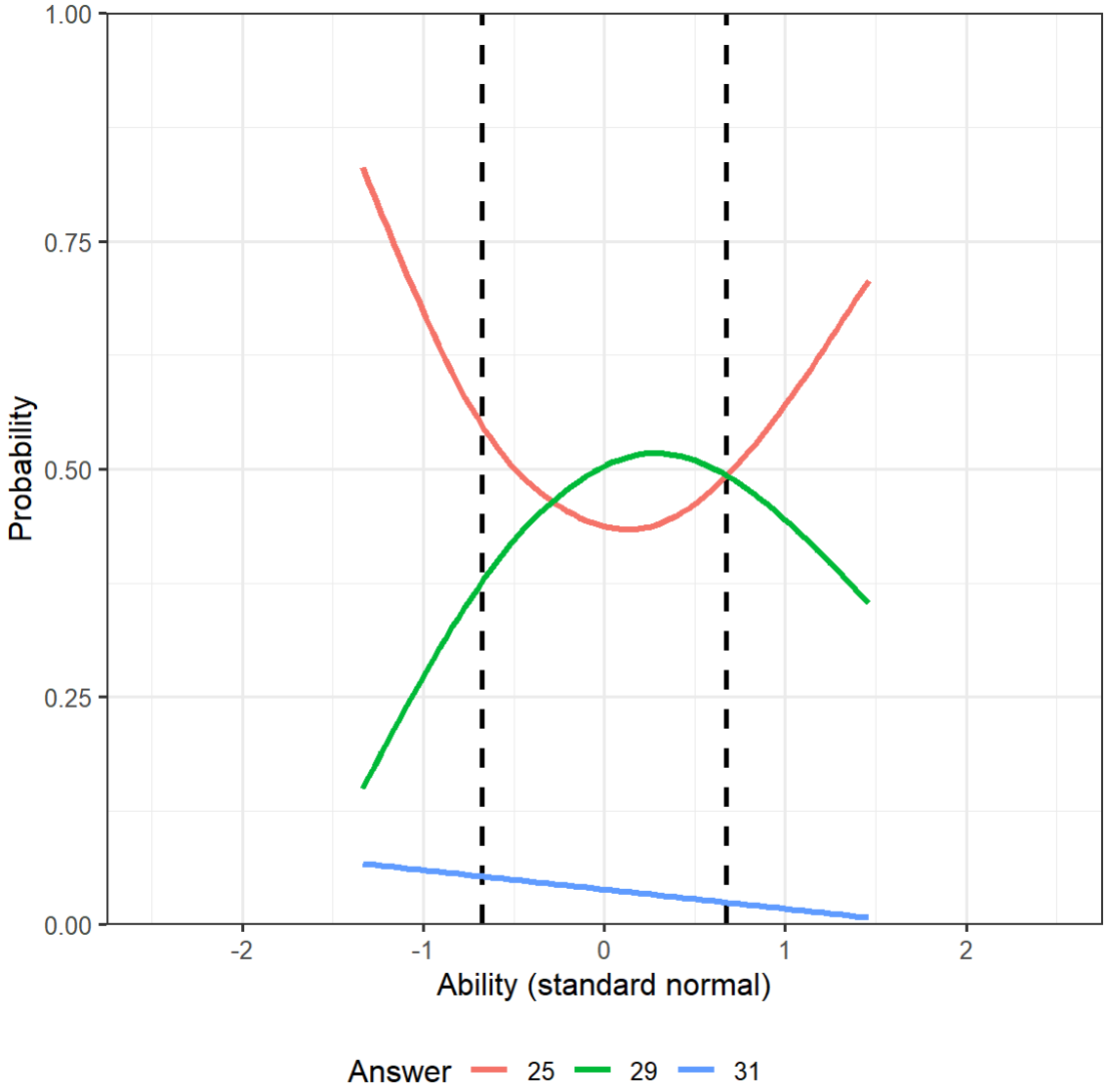
Answer	Frequency	Percent
24 (correct)	257	91.8%
25	13	4.6%
29	8	2.9%
31	2	0.7%
Overall	280	100.0%

10.1.32.4 Distractors functioning:

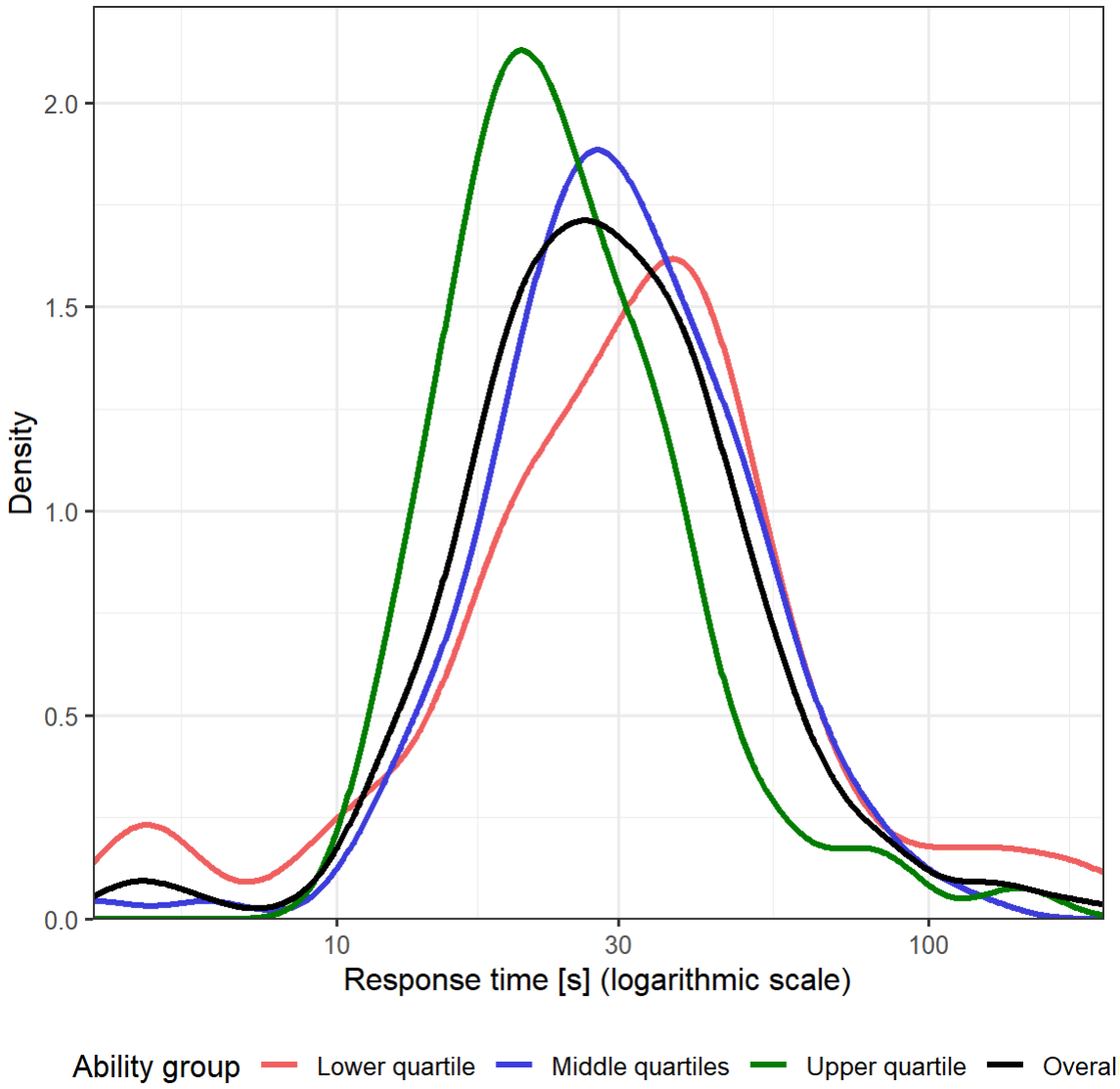
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
25	8	4	1	13	61.5%	44.4%	100.0%	56.5%
29	3	5	0	8	23.1%	55.6%	0.0%	34.8%
31	2	0	0	2	15.4%	0.0%	0.0%	8.7%
Overall	13	9	1	23	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



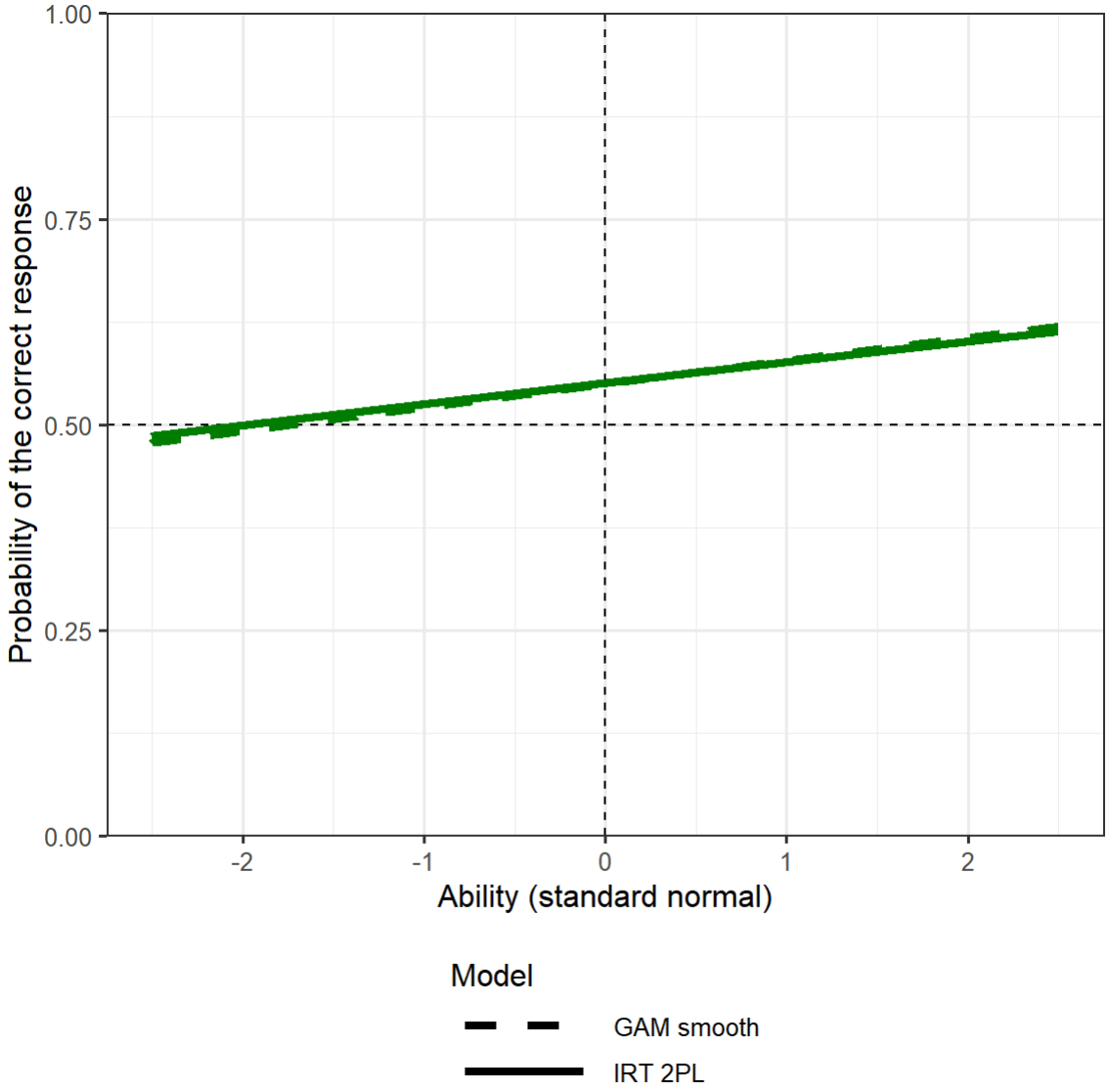
10.1.32.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.6	21.4	31.7	42.8	197.6	3.4	0.7
Middle quartiles	3.9	21.5	29.1	40.7	118.2	3.4	0.5
Upper quartile	11.4	17.8	23.0	31.5	142.6	3.2	0.5
Overall	3.9	20.2	28.0	39.6	197.6	3.3	0.6

10.1.33 M041264 (TIMSS)

10.1.33.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	54.93%	304	TIMSS	MC	0.06	0.05	0.10	−1.96	0.000

10.1.33.2 Gender DIF:

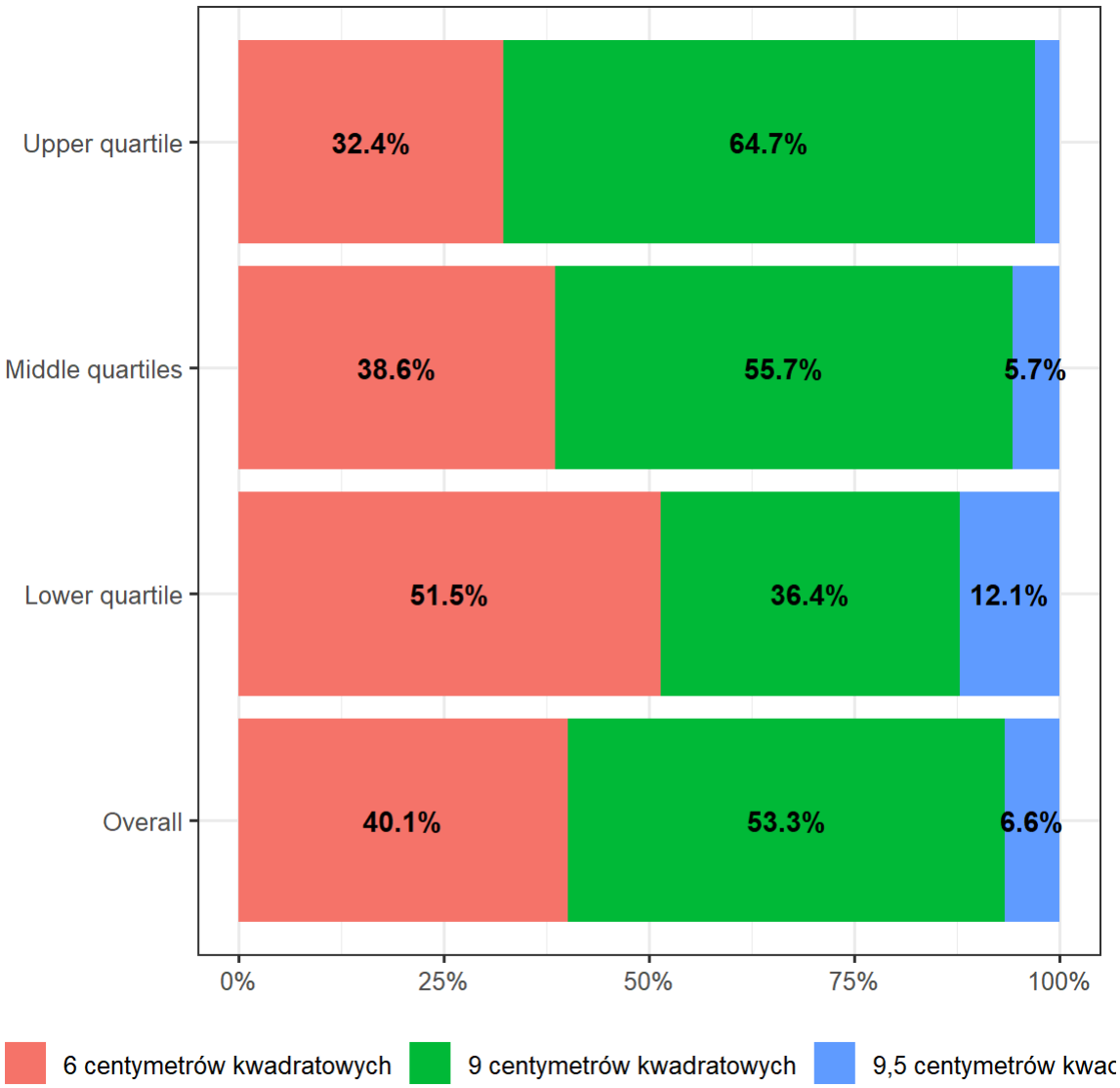
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.42	0.812	1.000	0.001	N/A

10.1.33.3 Frequency of answers:

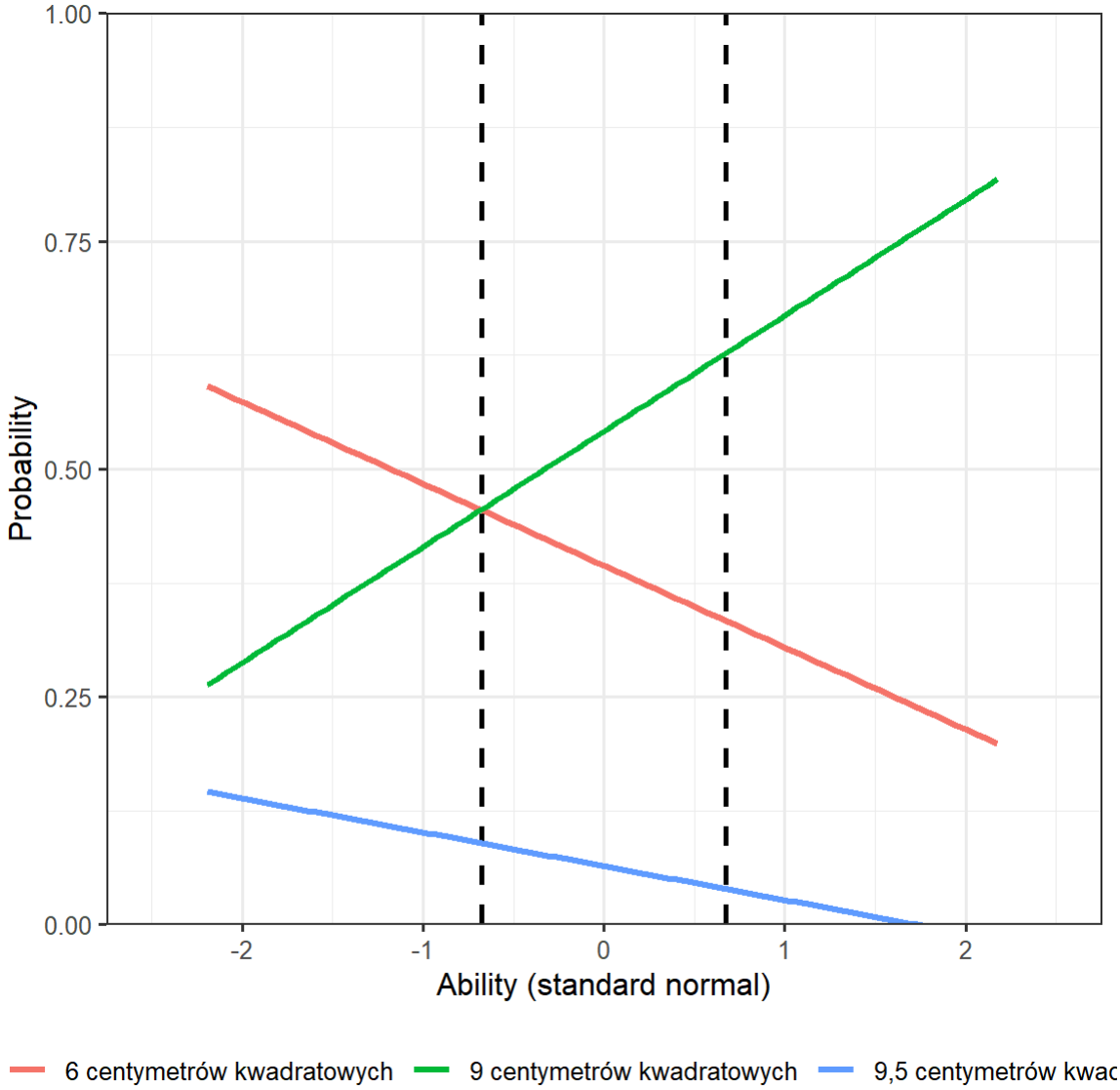
Answer	Frequency	Percent
4,5 centymetrów kwadratowych (correct)	167	54.9%
6 centymetrów kwadratowych	55	18.1%
9 centymetrów kwadratowych	73	24.0%
9,5 centymetrów kwadratowych	9	3.0%
Overall	304	100.0%

10.1.33.4 Distractors functioning:

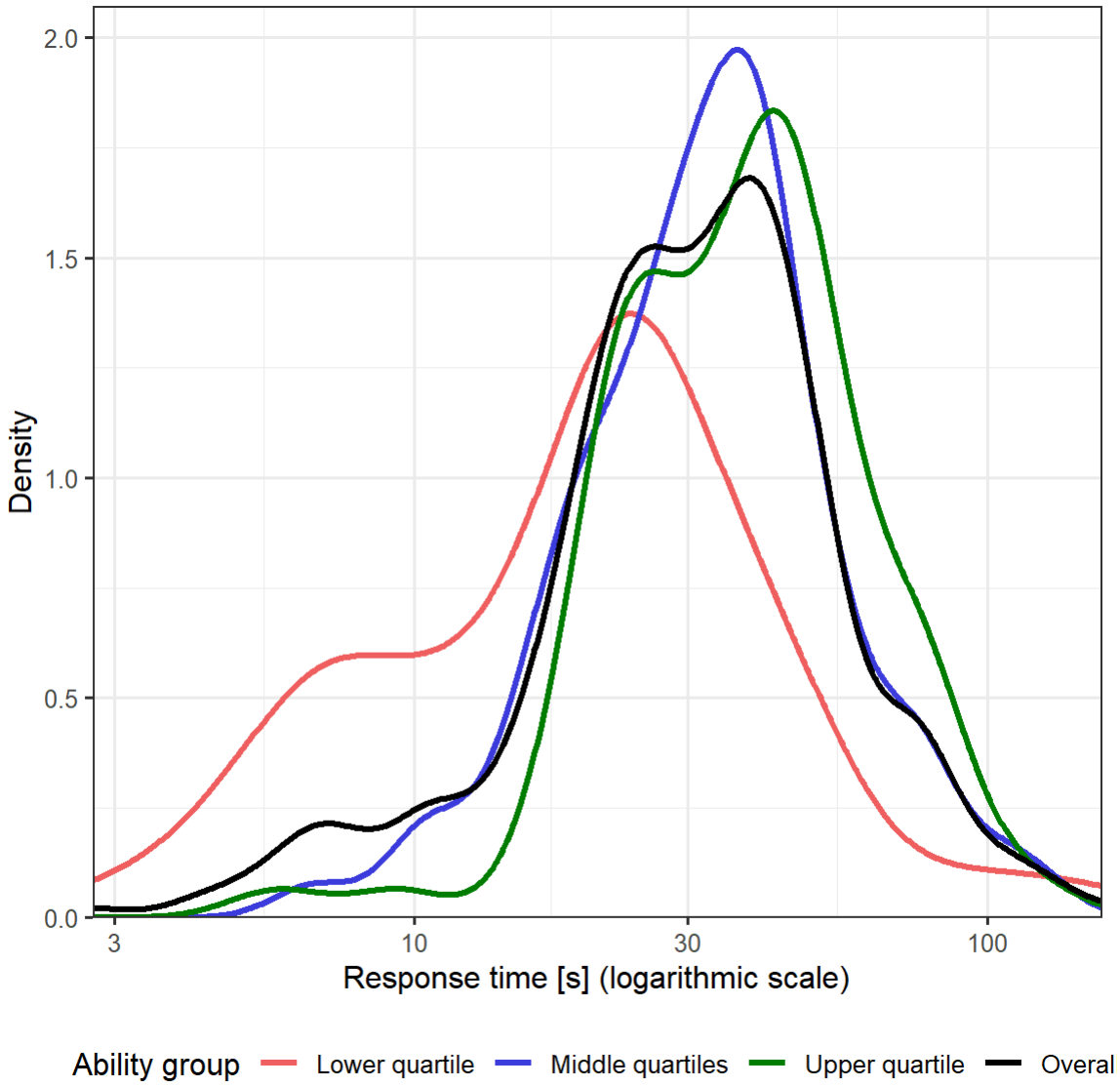
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
6 centymetrów kwadratowych	17	27	11	55	51.5%	38.6%	32.4%	40.1%
9 centymetrów kwadratowych	12	39	22	73	36.4%	55.7%	64.7%	53.3%
9,5 centymetrów kwadratowych	4	4	1	9	12.1%	5.7%	2.9%	6.6%
Overall	33	70	34	137	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



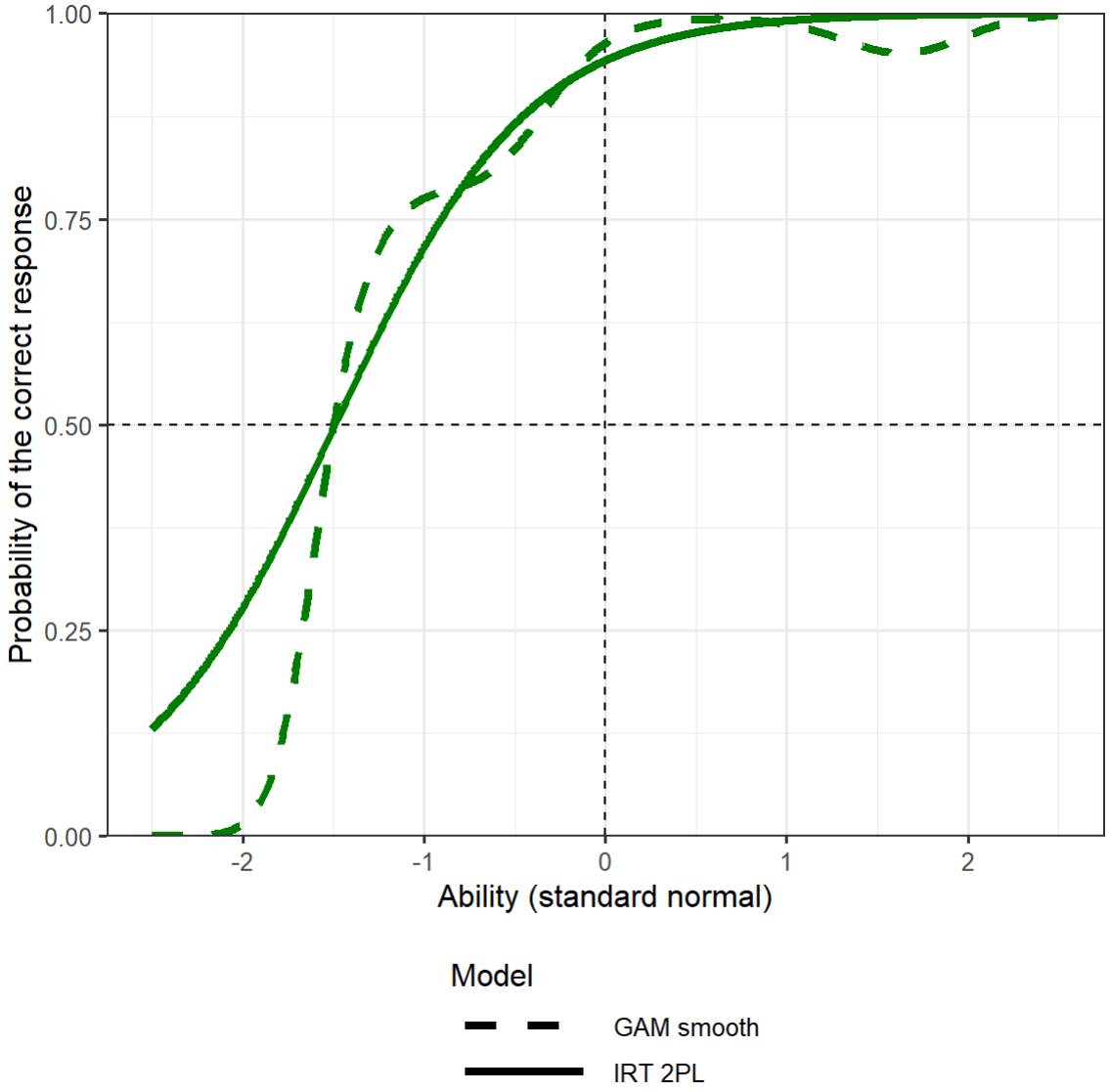
10.1.33.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.8	11.3	21.5	29.5	158.2	3.0	0.8
Middle quartiles	6.3	23.2	33.4	43.1	126.5	3.5	0.5
Upper quartile	5.8	25.2	37.8	50.1	123.6	3.6	0.5
Overall	2.8	21.8	31.2	43.5	158.2	3.4	0.6

10.1.34 M100037 (LLM)

10.1.34.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	86.84%	304	LLM simple setup	MC	0.49	0.31	1.88	−1.49	0.000

10.1.34.2 Gender DIF:

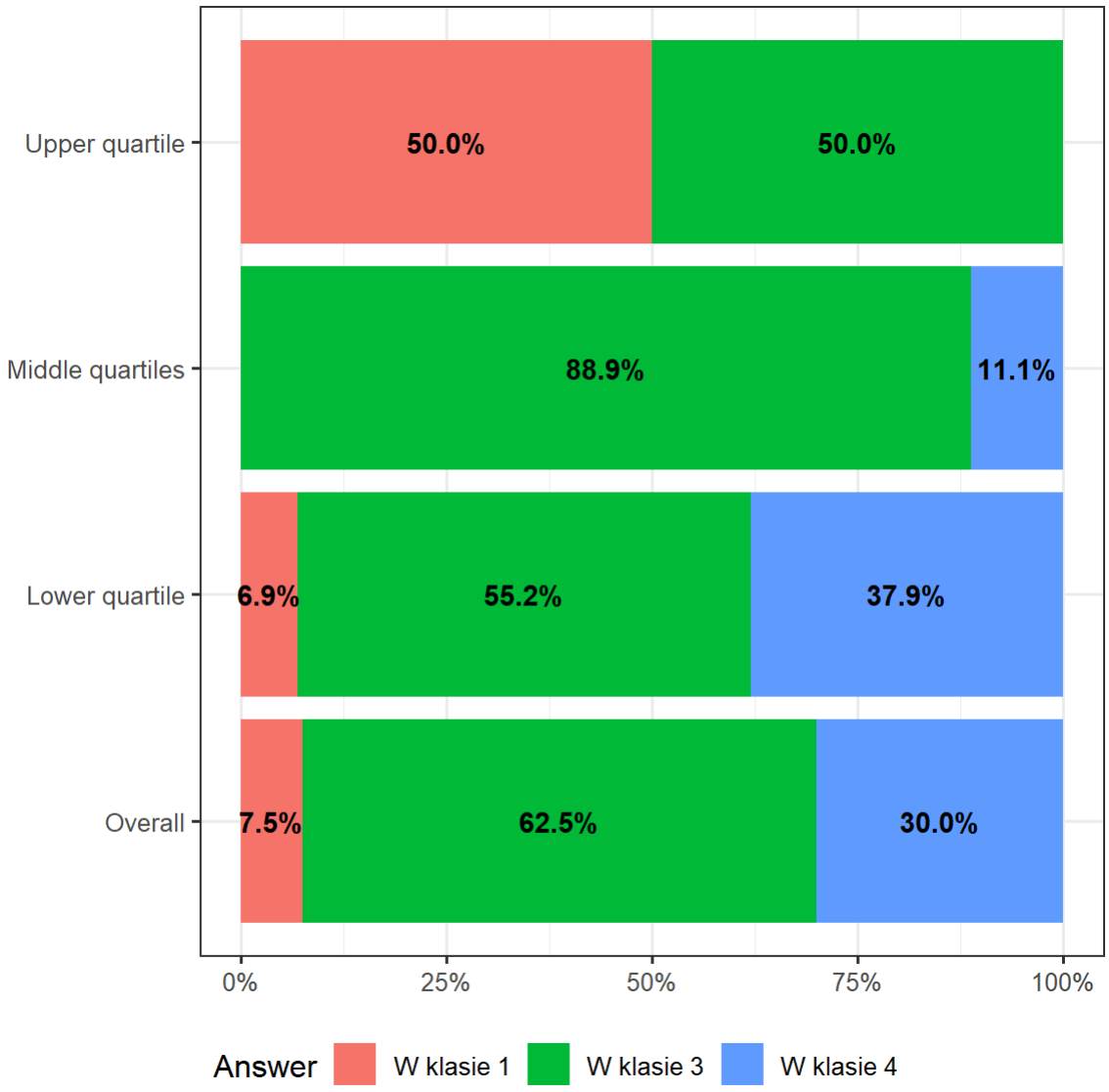
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.38	0.827	1.000	0.002	N/A

10.1.34.3 Frequency of answers:

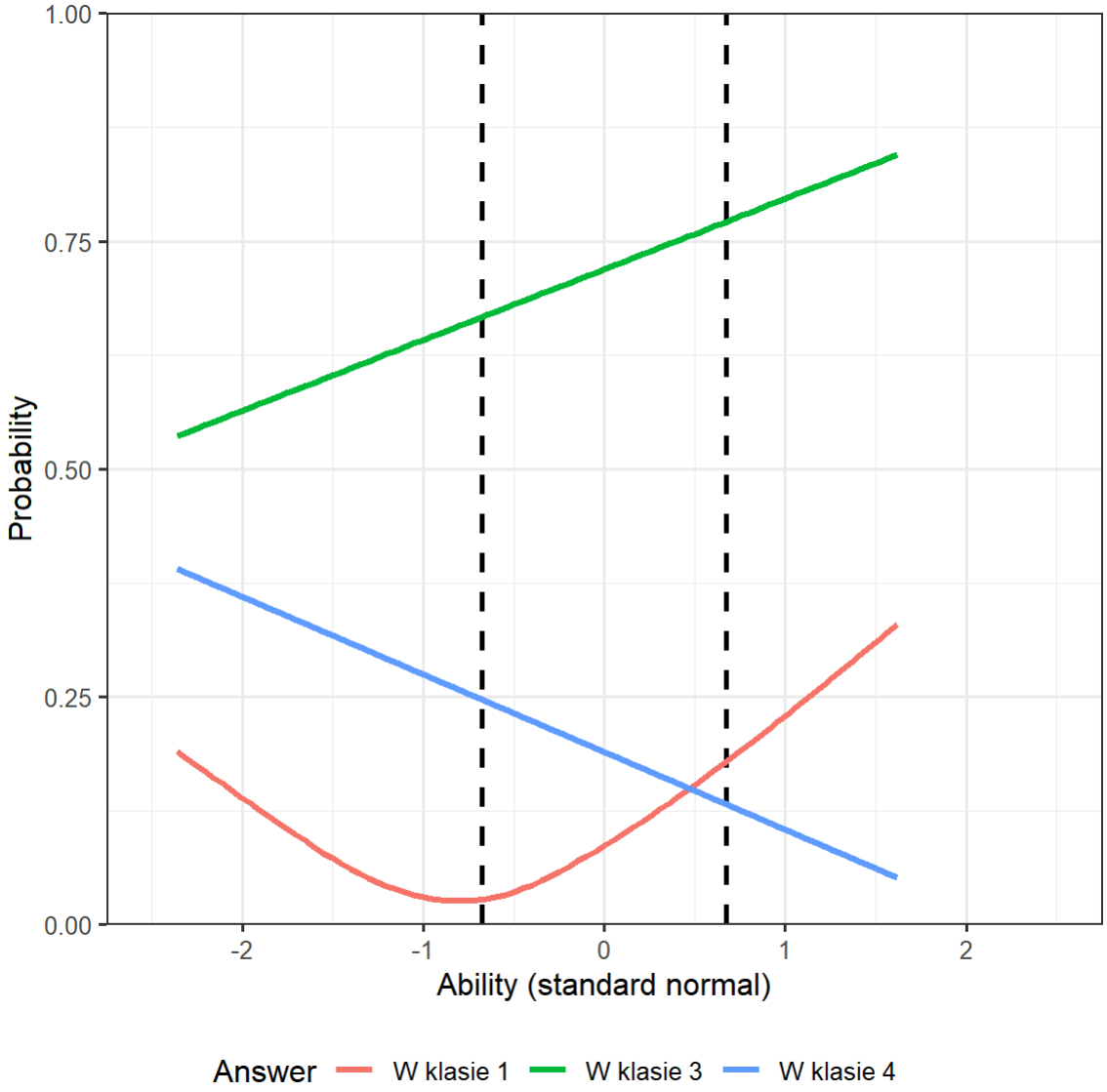
Answer	Frequency	Percent
W klasie 1	3	1.0%
W klasie 2 (correct)	264	86.8%
W klasie 3	25	8.2%
W klasie 4	12	3.9%
Overall	304	100.0%

10.1.34.4 Distractors functioning:

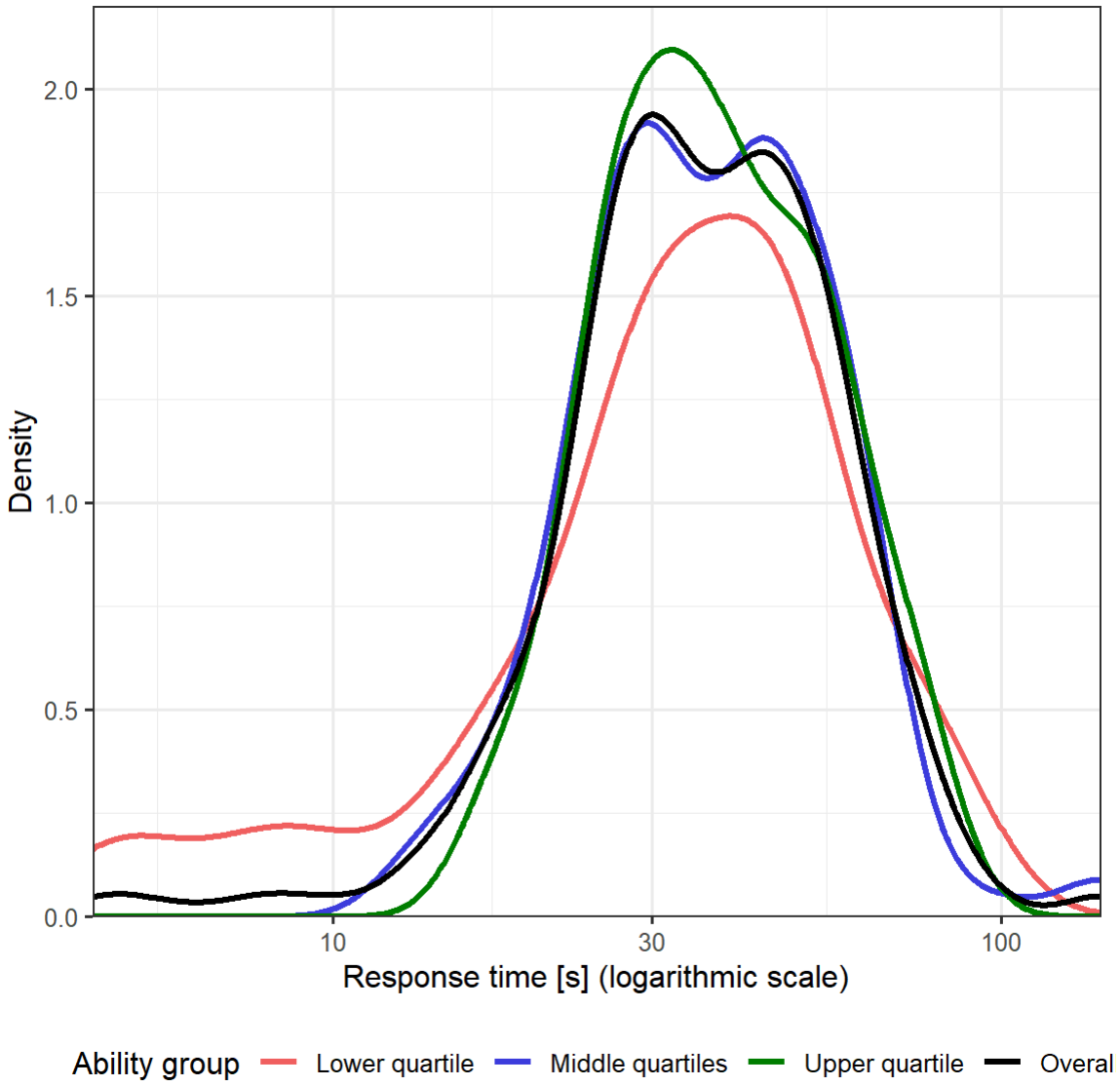
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
W klasie 1	2	0	1	3	6.9%	0.0%	50.0%	7.5%
W klasie 3	16	8	1	25	55.2%	88.9%	50.0%	62.5%
W klasie 4	11	1	0	12	37.9%	11.1%	0.0%	30.0%
Overall	29	9	2	40	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



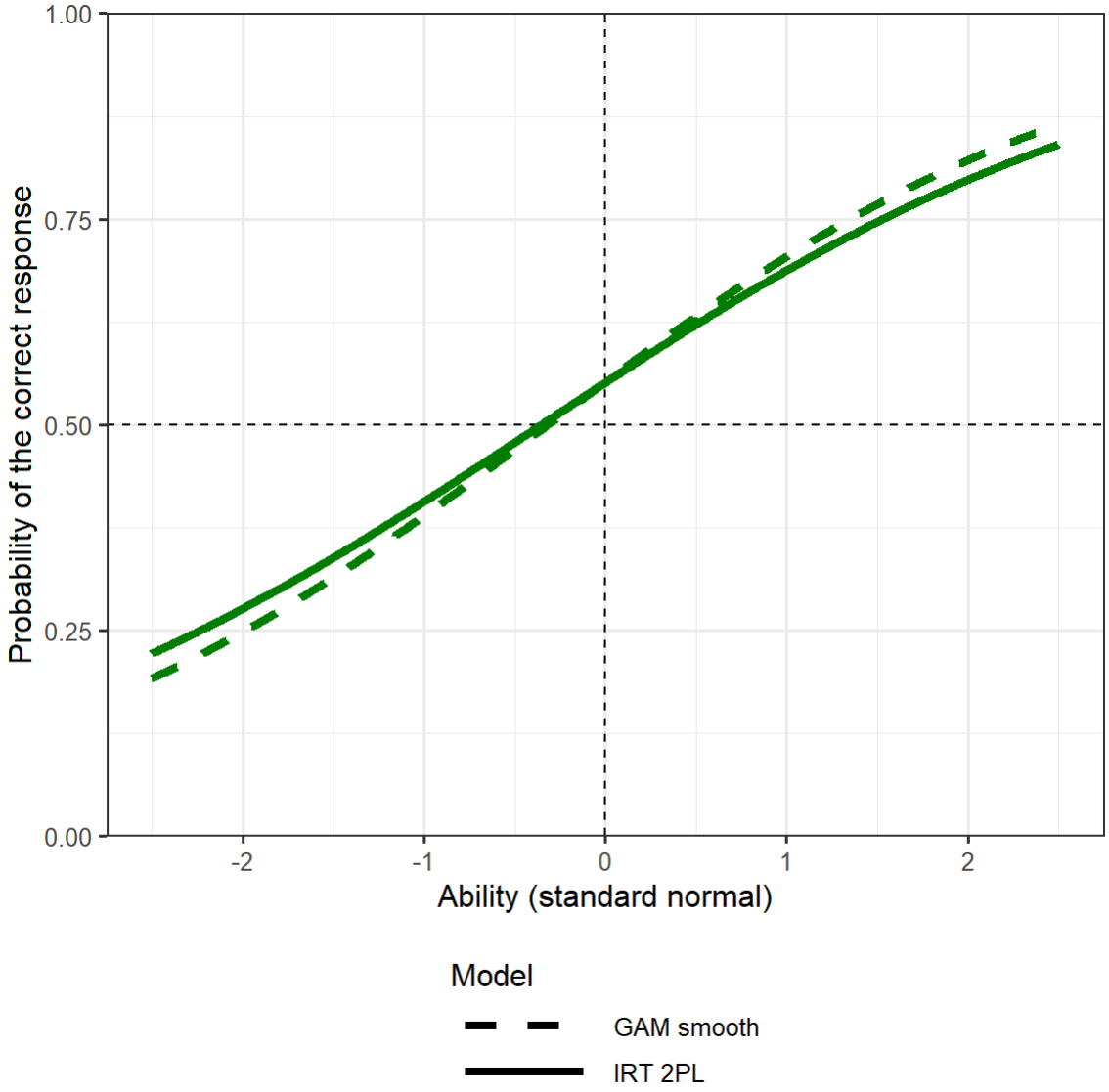
10.1.34.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.4	24.2	34.4	48.1	90.5	3.4	0.7
Middle quartiles	12.8	26.8	36.3	48.1	140.5	3.6	0.4
Upper quartile	16.9	28.1	36.7	52.1	81.1	3.6	0.4
Overall	4.4	27.2	35.3	48.7	140.5	3.6	0.5

10.1.35 M061206 (TIMSS)

10.1.35.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	54.10%	306	TIMSS	MC	0.30	0.30	0.58	−0.35	0.019

10.1.35.2 Gender DIF:

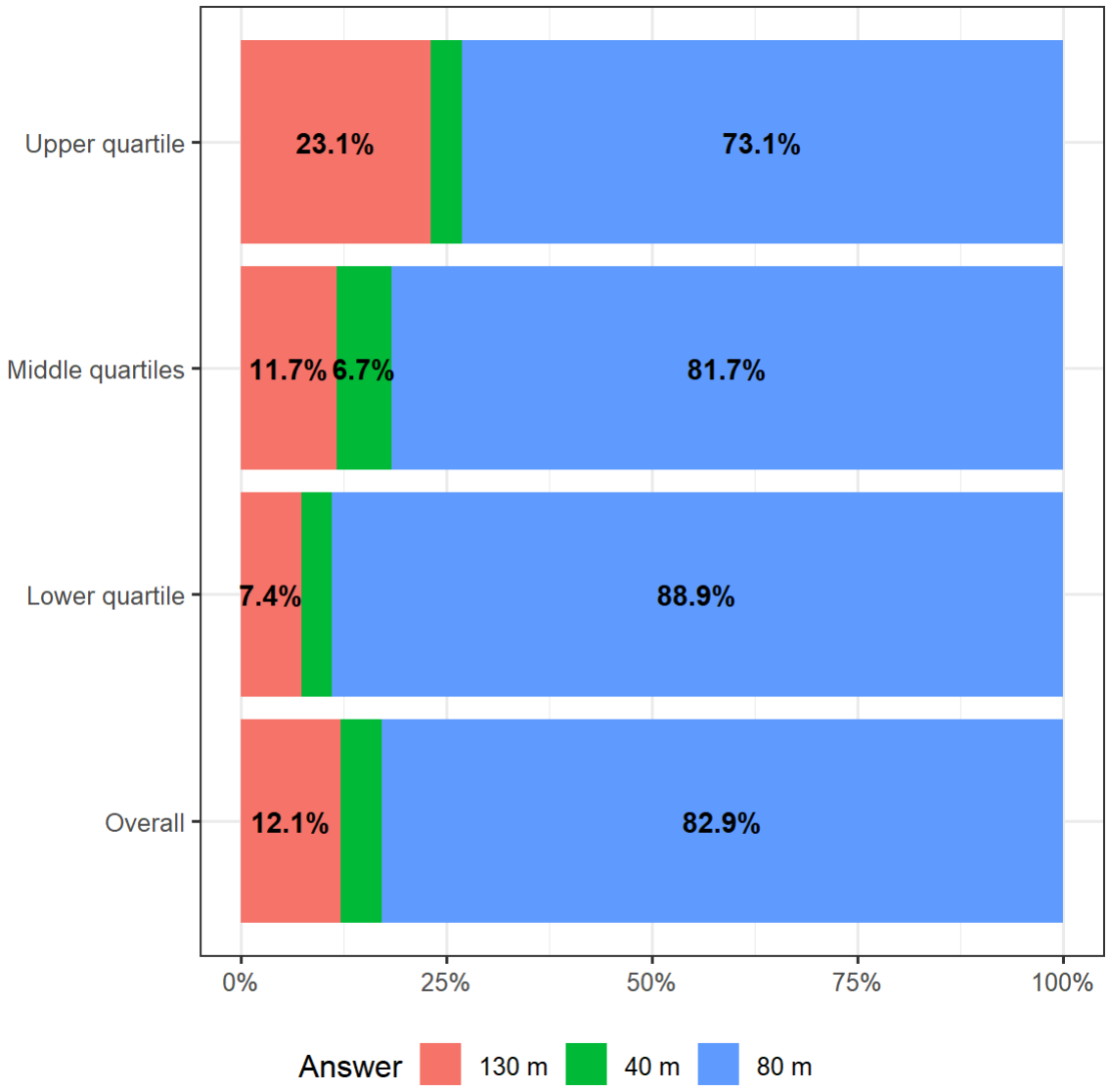
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.10	0.576	1.000	0.003	N/A

10.1.35.3 Frequency of answers:

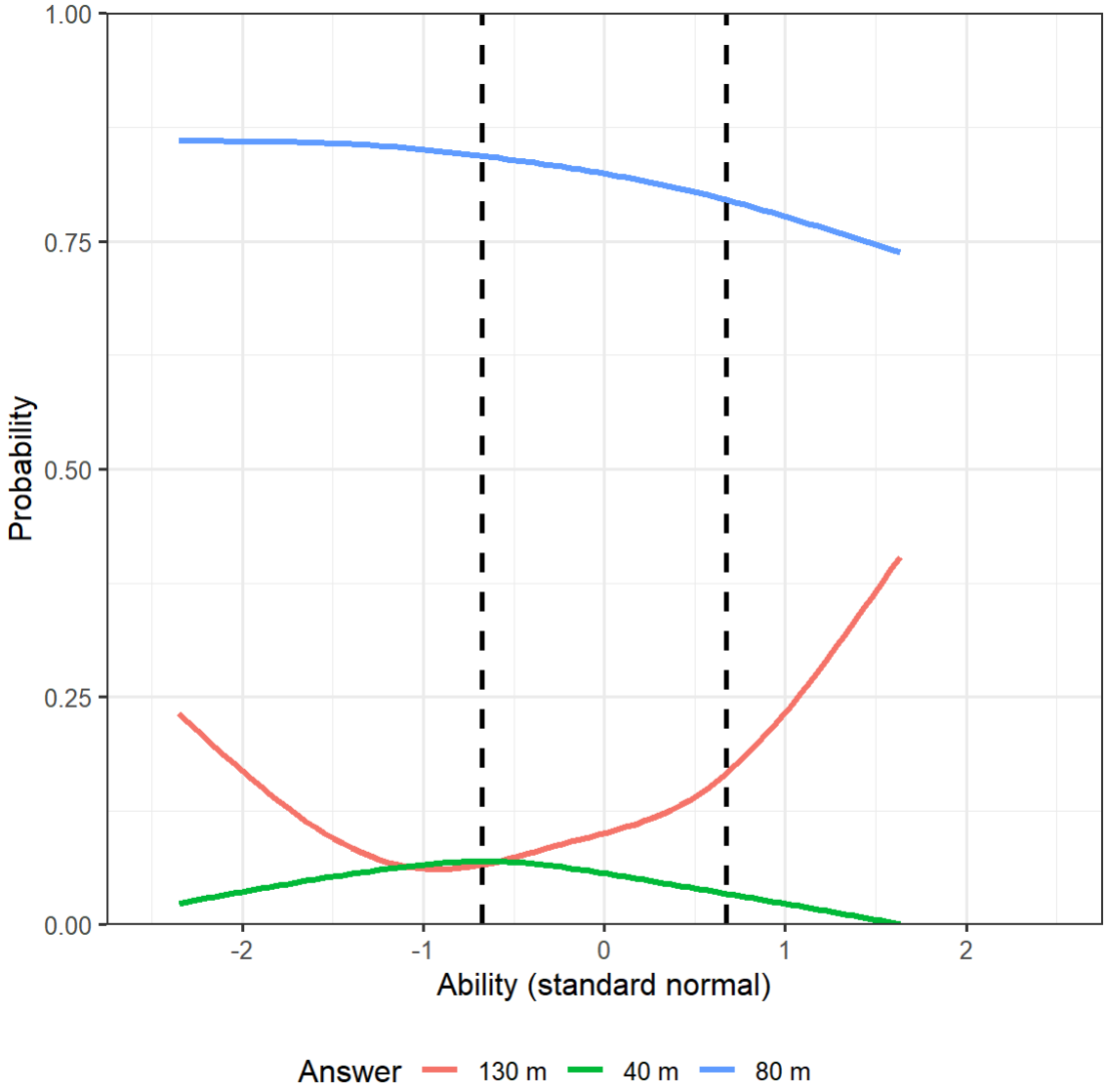
Answer	Frequency	Percent
100 m (correct)	165	54.1%
130 m	17	5.6%
40 m	7	2.3%
80 m	116	38.0%
Overall	305	100.0%

10.1.35.4 Distractors functioning:

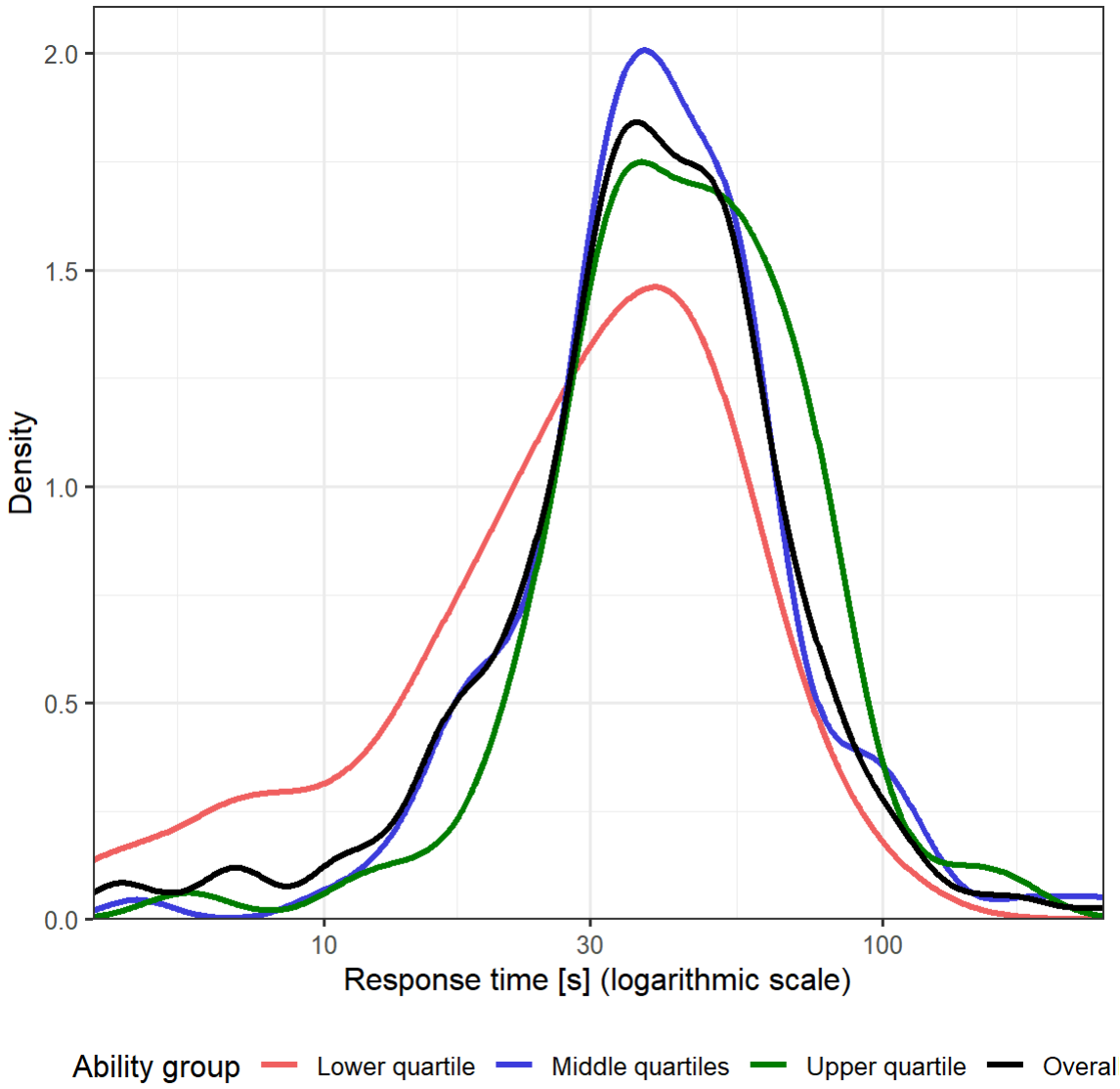
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
130 m	4	7	6	17	7.4%	11.7%	23.1%	12.1%
40 m	2	4	1	7	3.7%	6.7%	3.8%	5.0%
80 m	48	49	19	116	88.9%	81.7%	73.1%	82.9%
Overall	54	60	26	140	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



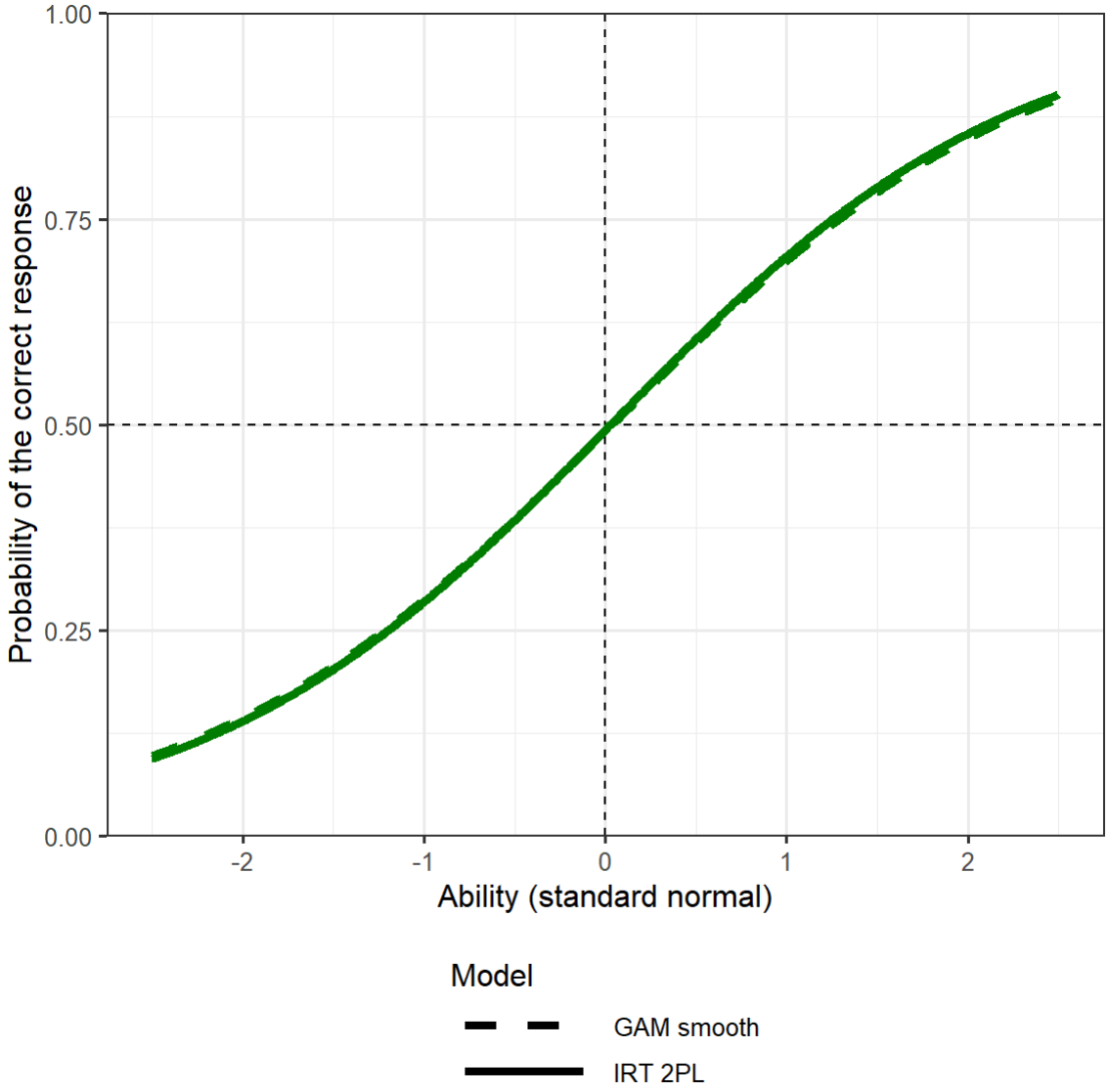
10.1.35.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.9	18.6	31.3	45.9	105.0	3.3	0.7
Middle quartiles	4.6	29.6	39.4	52.9	247.5	3.7	0.5
Upper quartile	5.7	32.1	45.5	61.7	168.5	3.8	0.5
Overall	3.9	28.7	37.7	52.7	247.5	3.6	0.6

10.1.36 M041034 (TIMSS)

10.1.36.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	48.52%	306	TIMSS	MC	0.38	0.44	0.90	0.03	0.011

10.1.36.2 Gender DIF:

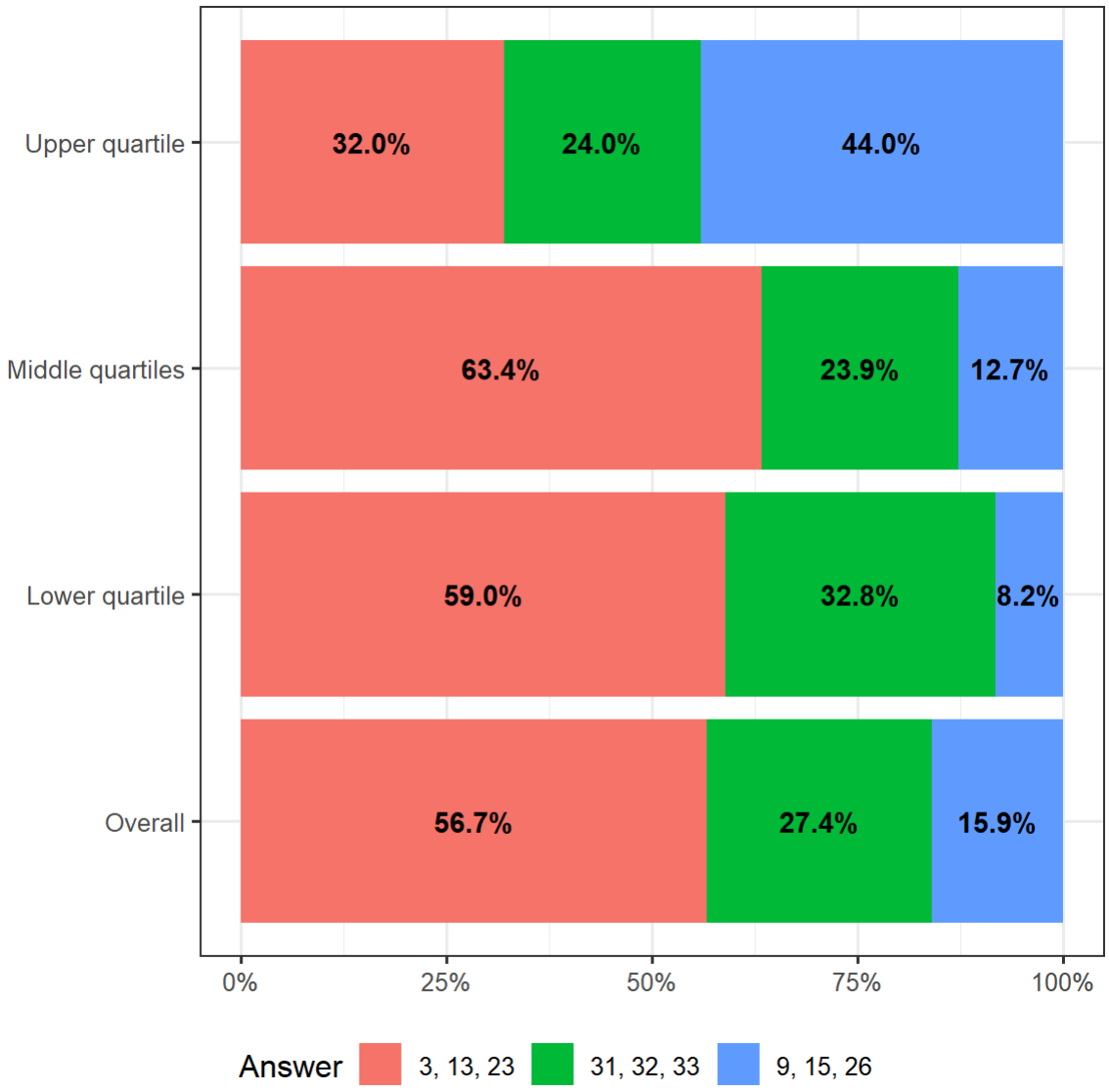
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.16	0.560	1.000	0.003	N/A

10.1.36.3 Frequency of answers:

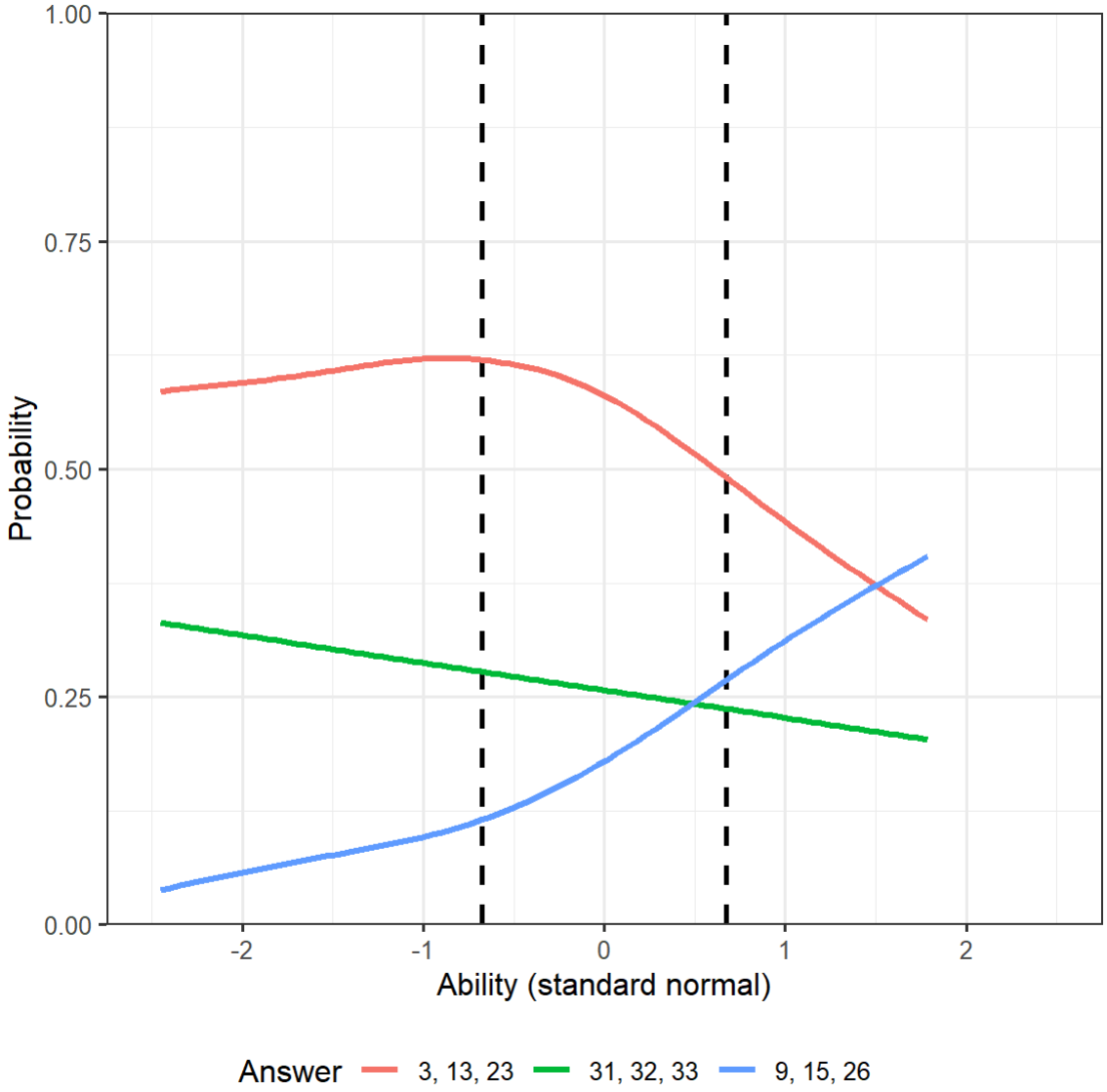
Answer	Frequency	Percent
3, 13, 23	89	29.2%
31, 32, 33	43	14.1%
6, 15, 21 (correct)	148	48.5%
9, 15, 26	25	8.2%
Overall	305	100.0%

10.1.36.4 Distractors functioning:

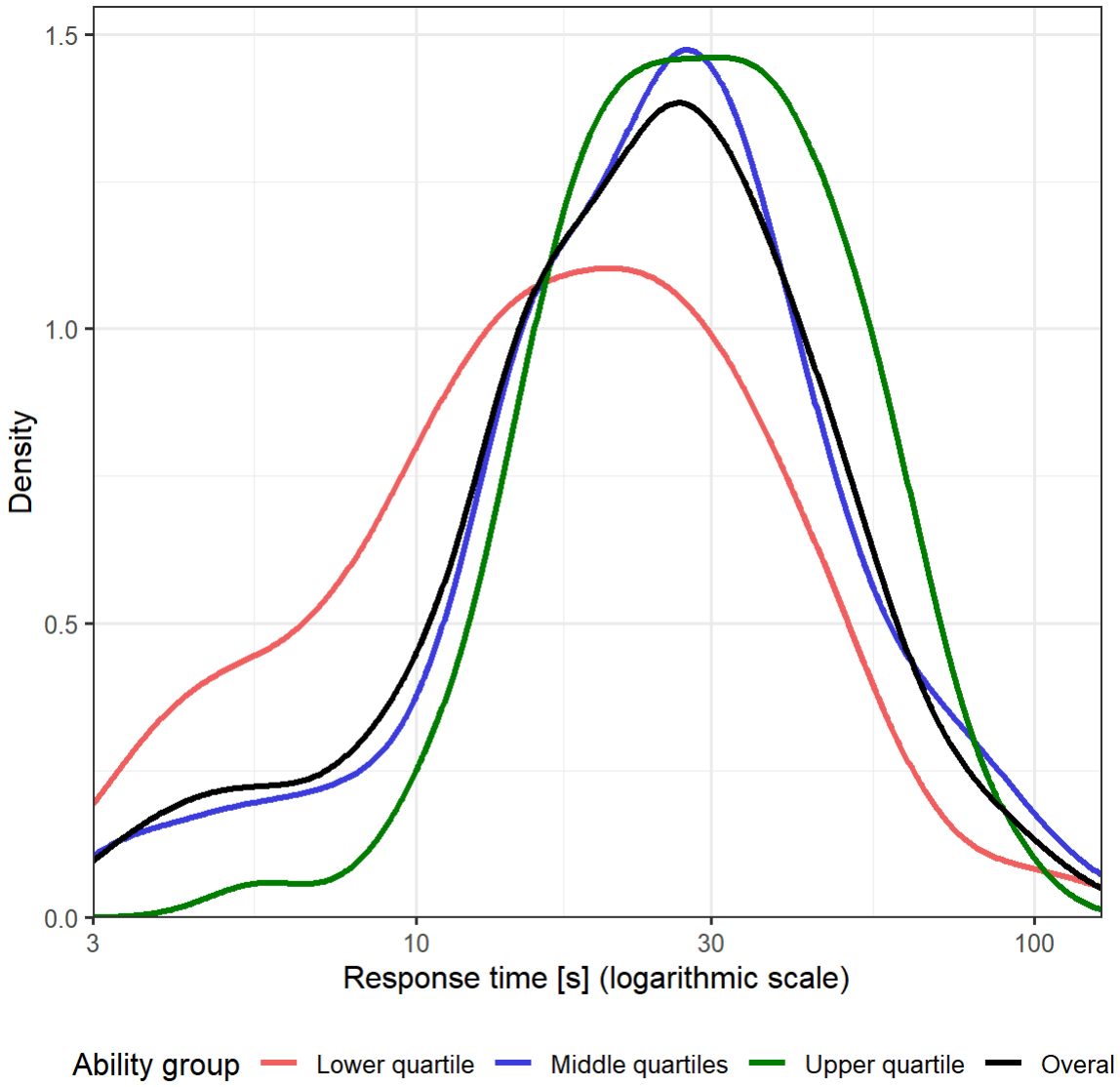
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
3, 13, 23	36	45	8	89	59.0%	63.4%	32.0%	56.7%
31, 32, 33	20	17	6	43	32.8%	23.9%	24.0%	27.4%
9, 15, 26	5	9	11	25	8.2%	12.7%	44.0%	15.9%
Overall	61	71	25	157	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



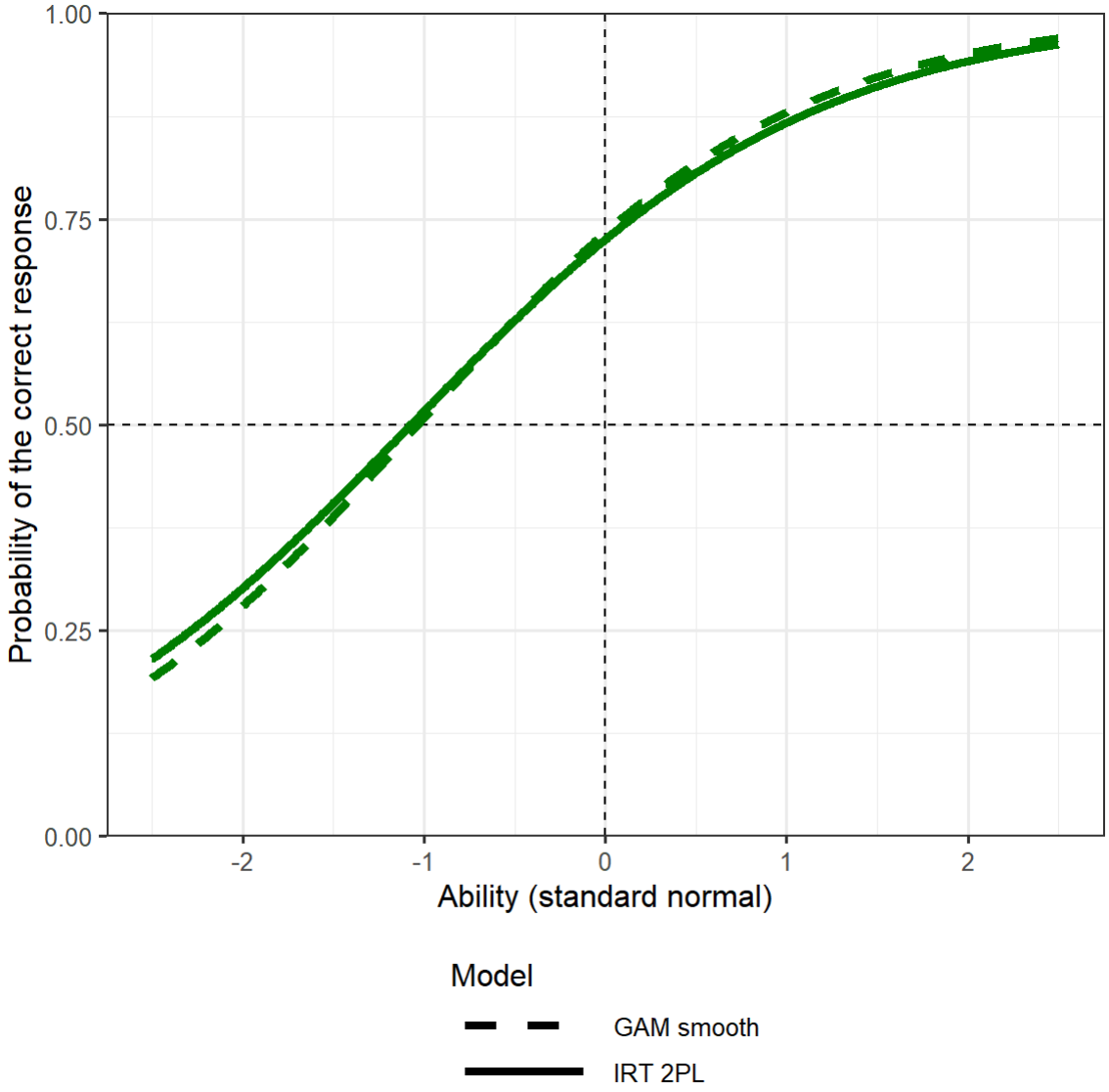
10.1.36.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	10.1	17.8	28.8	111.4	2.8	0.8
Middle quartiles	3.0	15.8	25.0	35.4	128.2	3.2	0.7
Upper quartile	5.6	19.4	29.2	42.8	88.3	3.3	0.5
Overall	3.0	15.0	24.2	36.3	128.2	3.1	0.7

10.1.37 M051129 (TIMSS)

10.1.37.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	70.41%	294	TIMSS	MC	0.39	0.31	0.91	−1.07	0.000

10.1.37.2 Gender DIF:

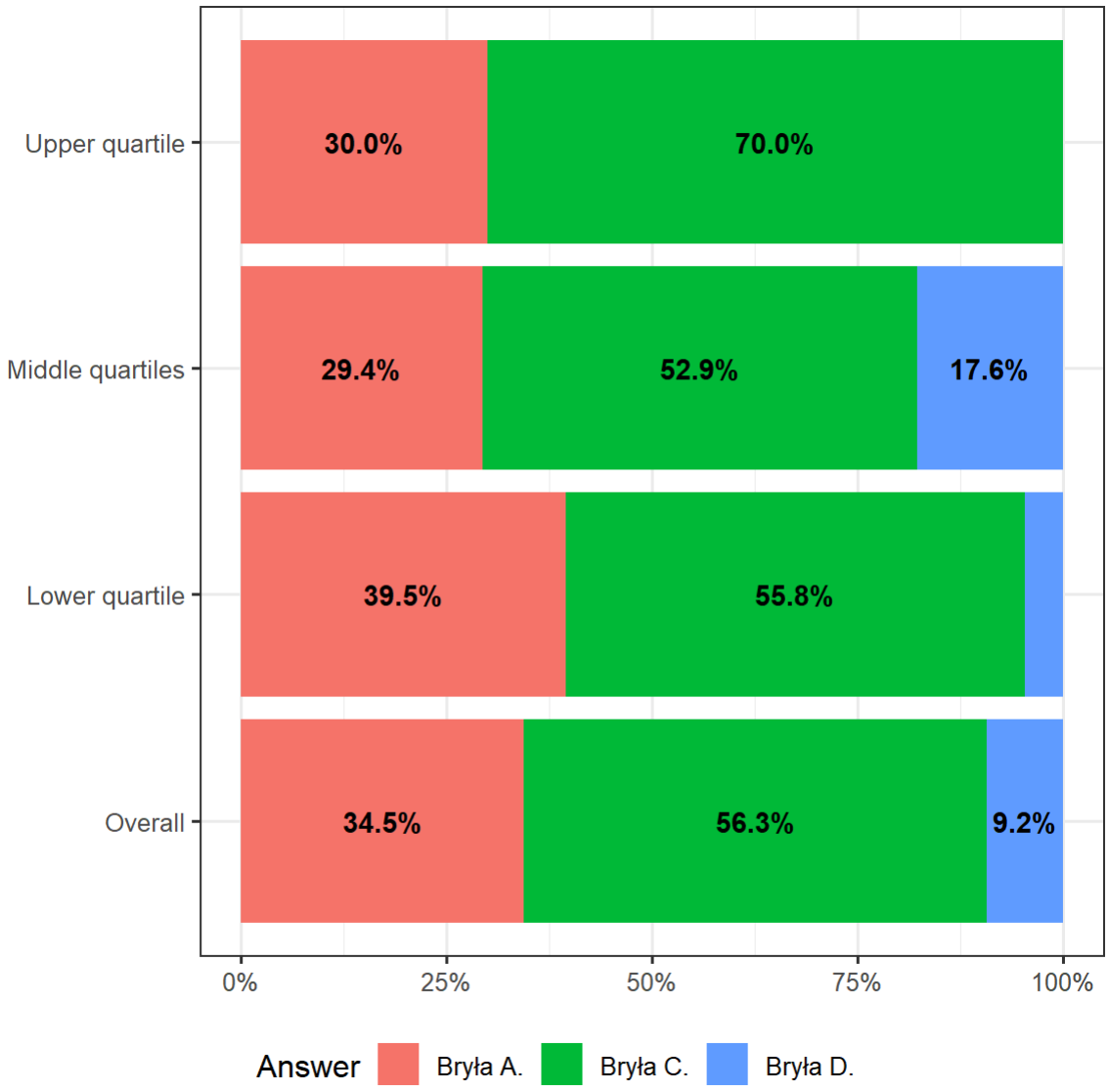
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.52	0.769	1.000	0.002	N/A

10.1.37.3 Frequency of answers:

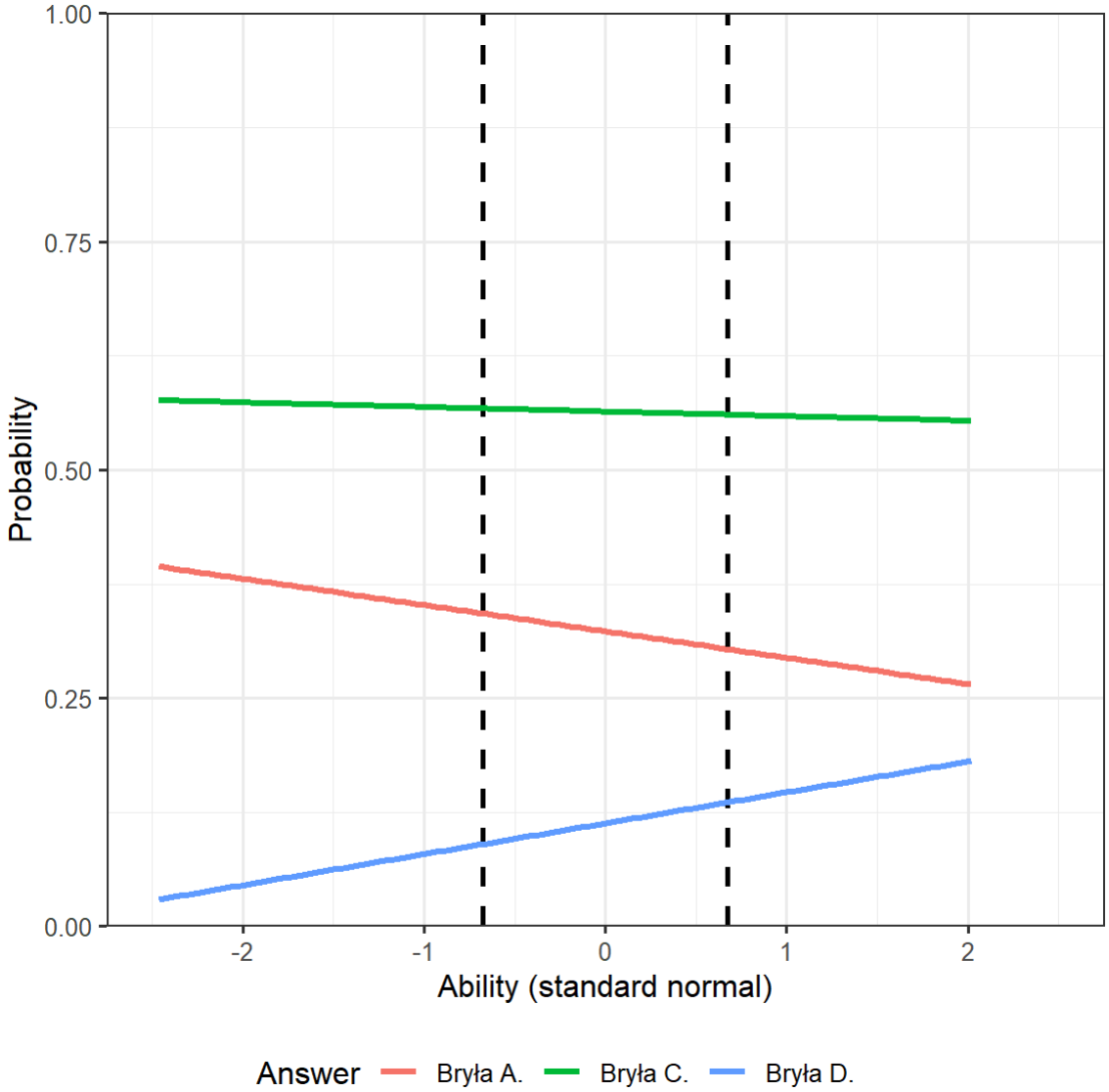
Answer	Frequency	Percent
Bryła A.	30	10.2%
Bryła B. (correct)	207	70.4%
Bryła C.	49	16.7%
Bryła D.	8	2.7%
Overall	294	100.0%

10.1.37.4 Distractors functioning:

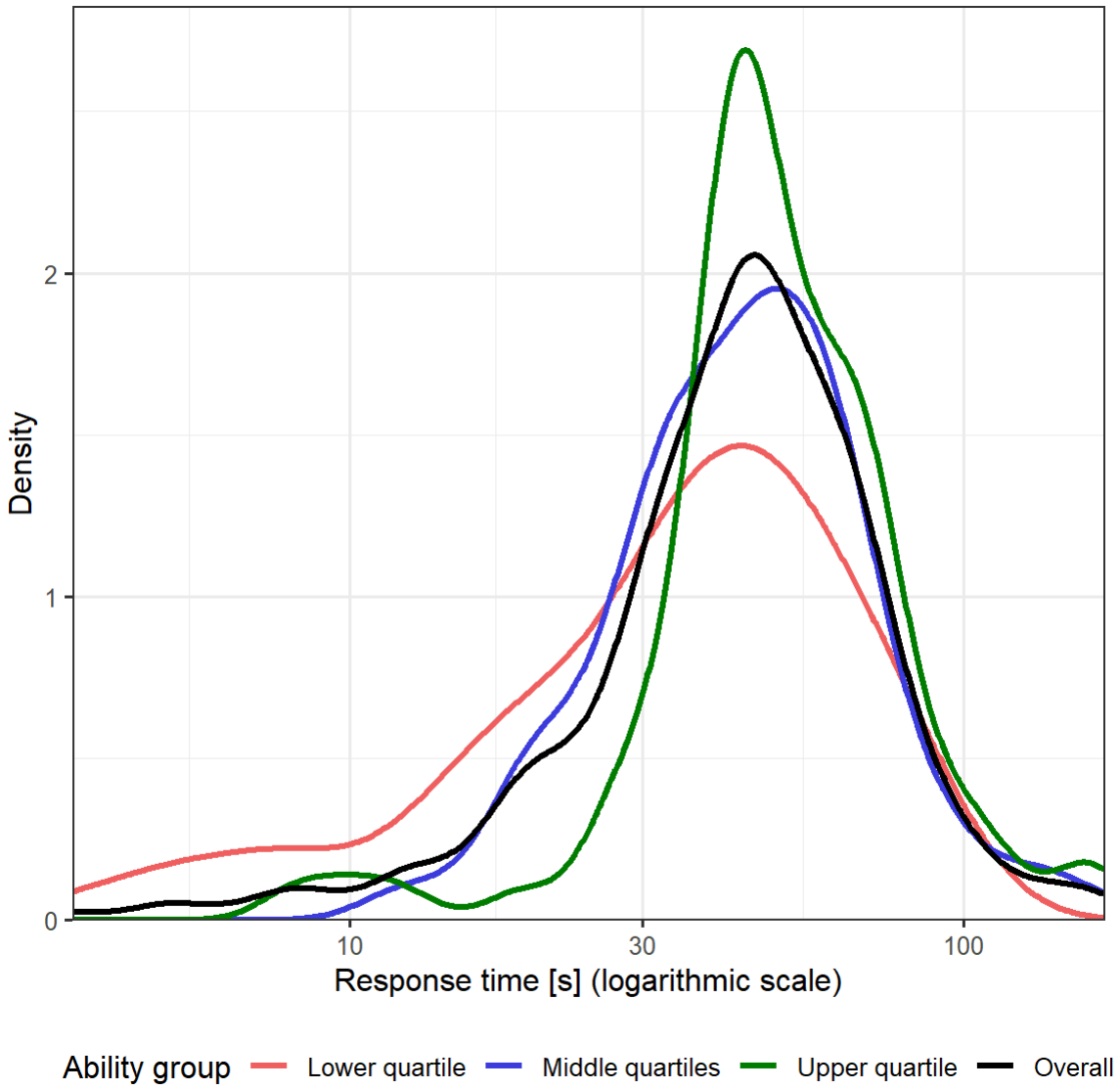
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Bryła A.	17	10	3	30	39.5%	29.4%	30.0%	34.5%
Bryła C.	24	18	7	49	55.8%	52.9%	70.0%	56.3%
Bryła D.	2	6	0	8	4.7%	17.6%	0.0%	9.2%
Overall	43	34	10	87	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



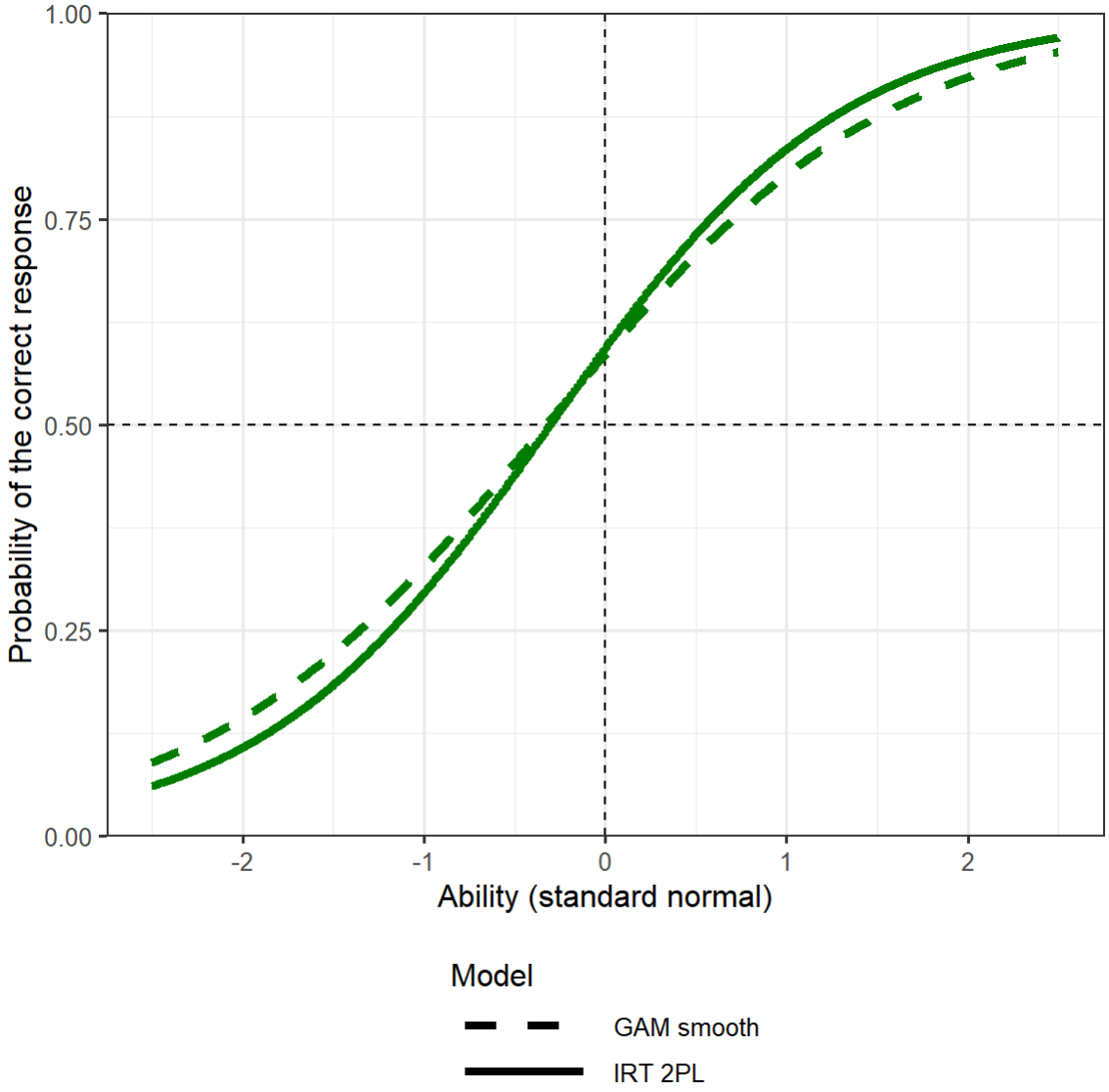
10.1.37.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	22.6	36.2	51.4	99.3	3.5	0.7
Middle quartiles	11.9	32.0	44.6	59.7	169.5	3.8	0.5
Upper quartile	8.1	40.5	47.6	64.8	162.8	3.9	0.5
Overall	3.5	32.0	44.3	59.8	169.5	3.7	0.6

10.1.38 M051077 (TIMSS)

10.1.38.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	56.07%	306	TIMSS	MC	0.44	0.47	1.25	−0.30	0.010

10.1.38.2 Gender DIF:

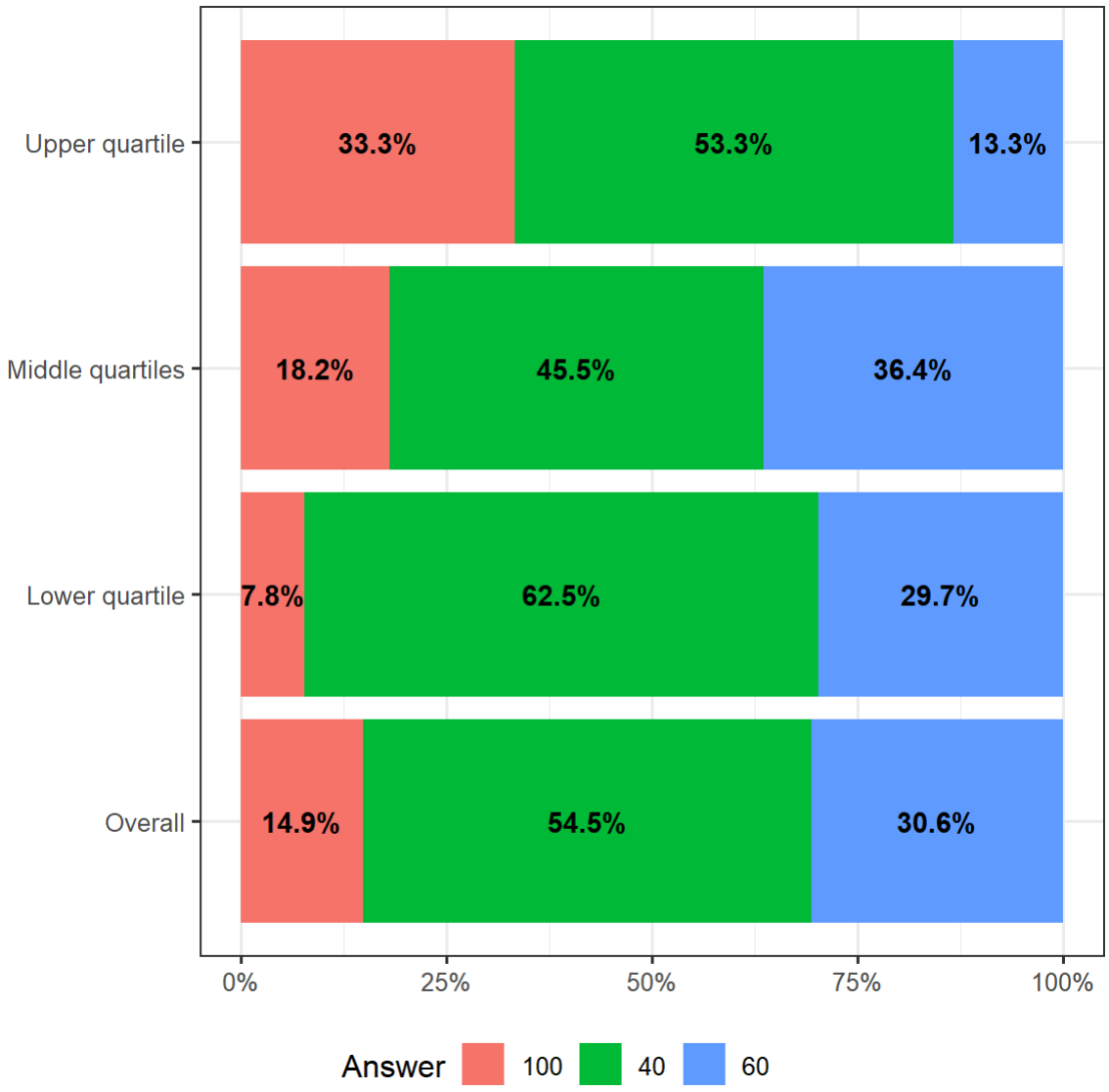
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.02	0.221	1.000	0.008	N/A

10.1.38.3 Frequency of answers:

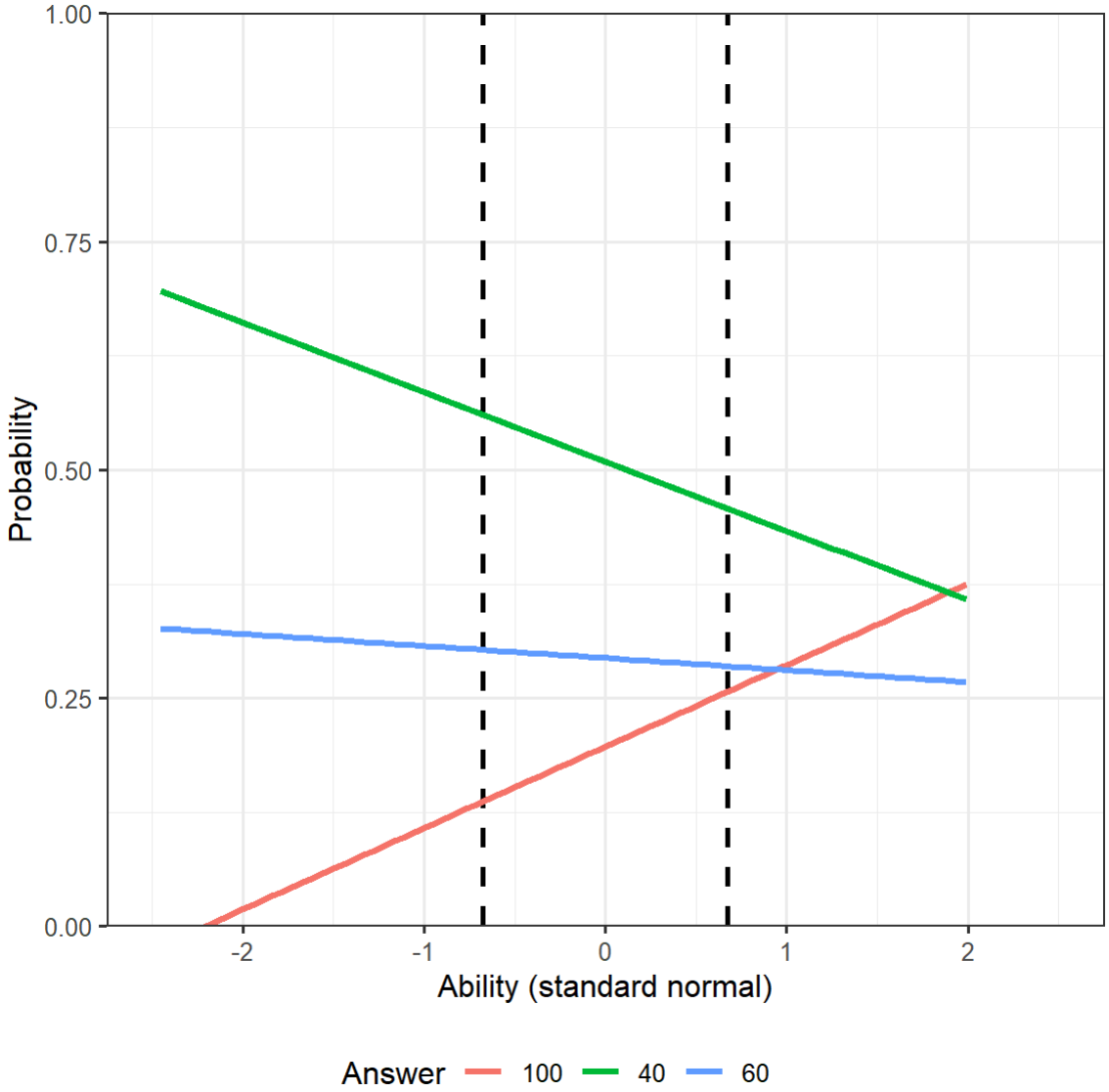
Answer	Frequency	Percent
40	73	23.9%
60	41	13.4%
80 (correct)	171	56.1%
100	20	6.6%
Overall	305	100.0%

10.1.38.4 Distractors functioning:

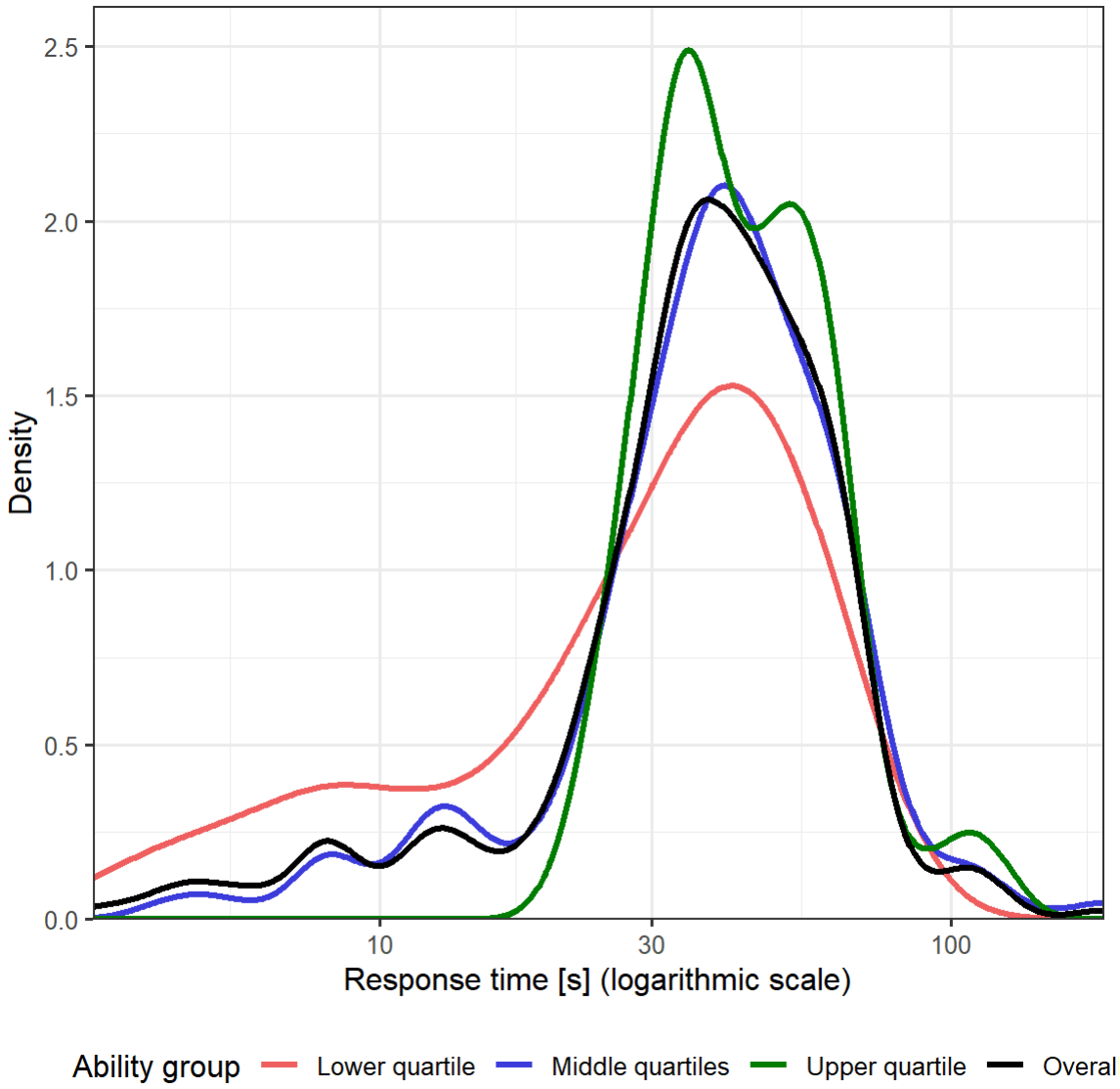
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
100	5	10	5	20	7.8%	18.2%	33.3%	14.9%
40	40	25	8	73	62.5%	45.5%	53.3%	54.5%
60	19	20	2	41	29.7%	36.4%	13.3%	30.6%
Overall	64	55	15	134	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



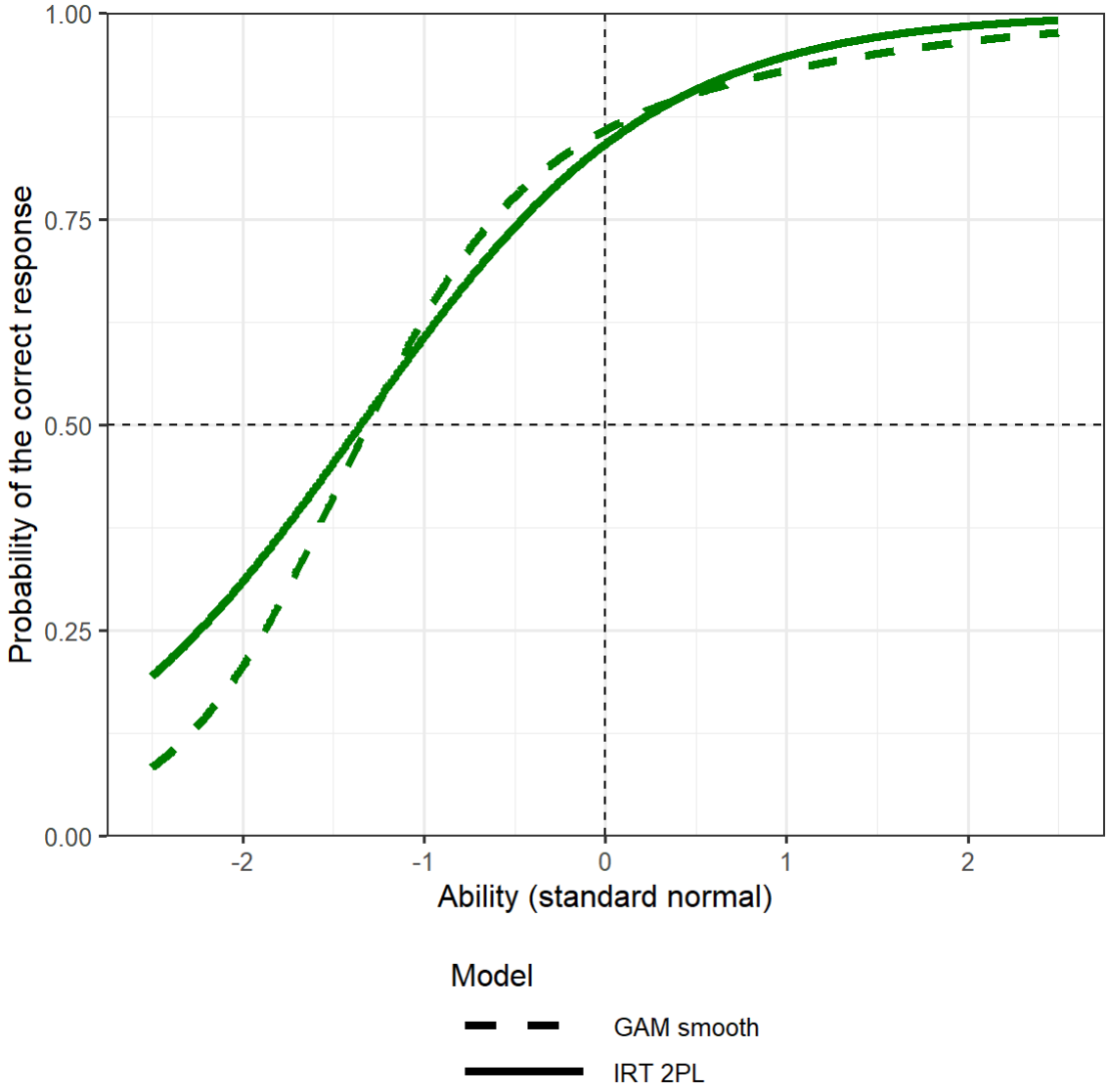
10.1.38.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	19.0	33.7	46.4	72.9	3.3	0.8
Middle quartiles	4.3	30.0	39.8	53.3	185.0	3.6	0.6
Upper quartile	22.6	33.0	40.2	55.0	117.3	3.8	0.4
Overall	3.1	29.0	38.6	51.2	185.0	3.6	0.6

10.1.39 M100008 (LLM)

10.1.39.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	78.19%	299	LLM simple setup	MC	0.43	0.27	1.23	−1.35	0.031

10.1.39.2 Gender DIF:

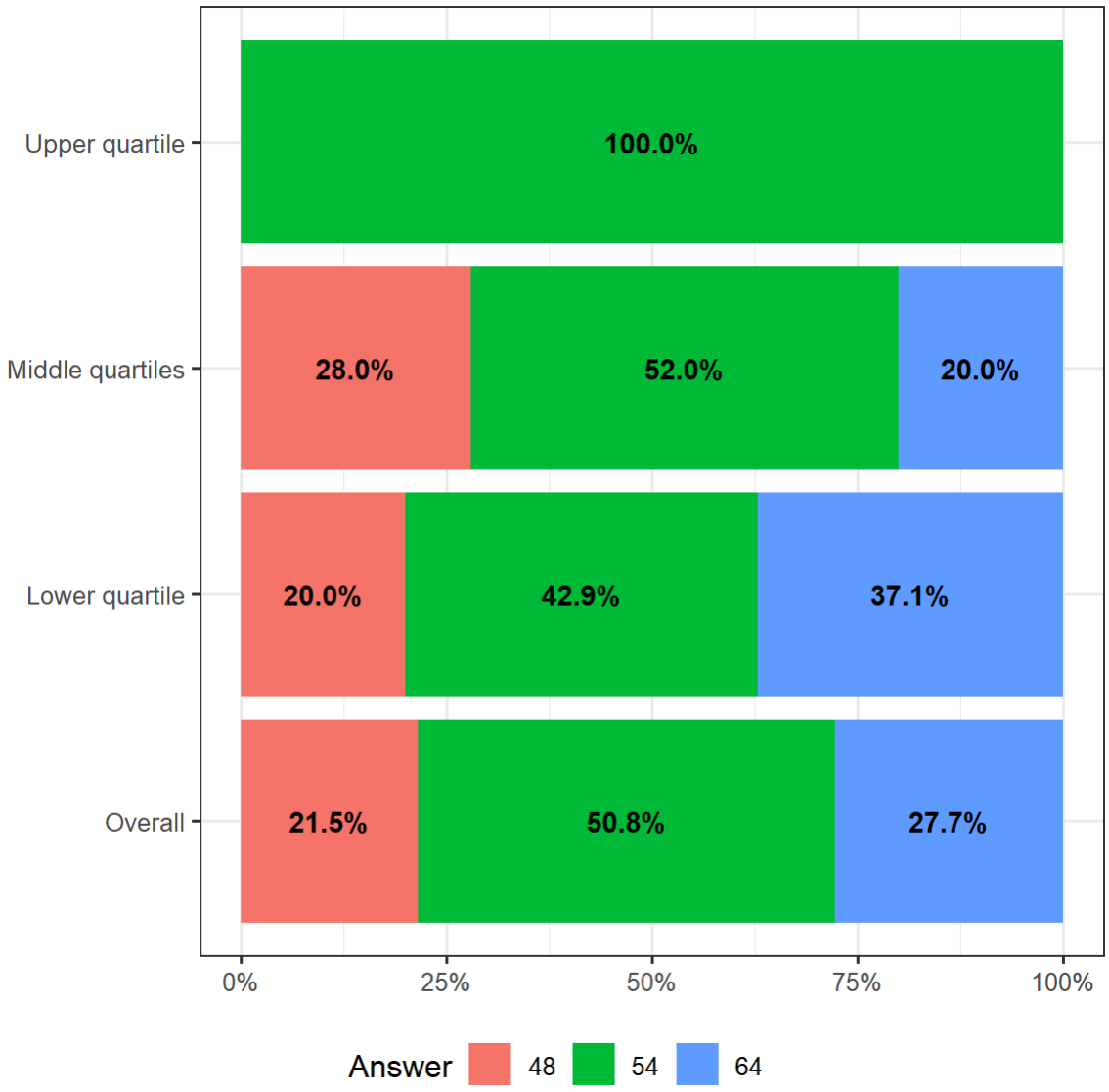
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.96	0.618	1.000	0.003	N/A

10.1.39.3 Frequency of answers:

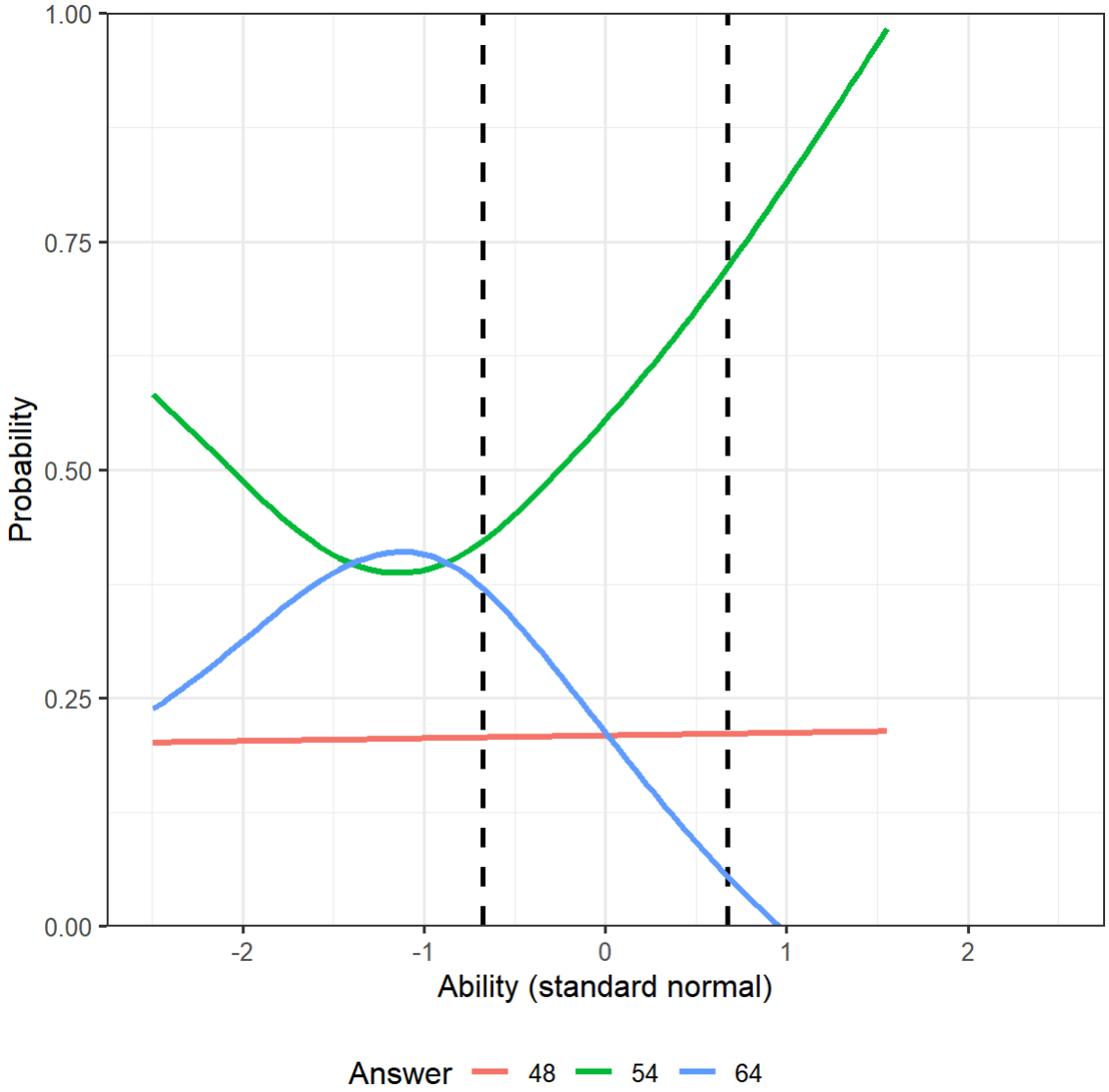
Answer	Frequency	Percent
48	14	4.7%
54	33	11.1%
56 (correct)	233	78.2%
64	18	6.0%
Overall	298	100.0%

10.1.39.4 Distractors functioning:

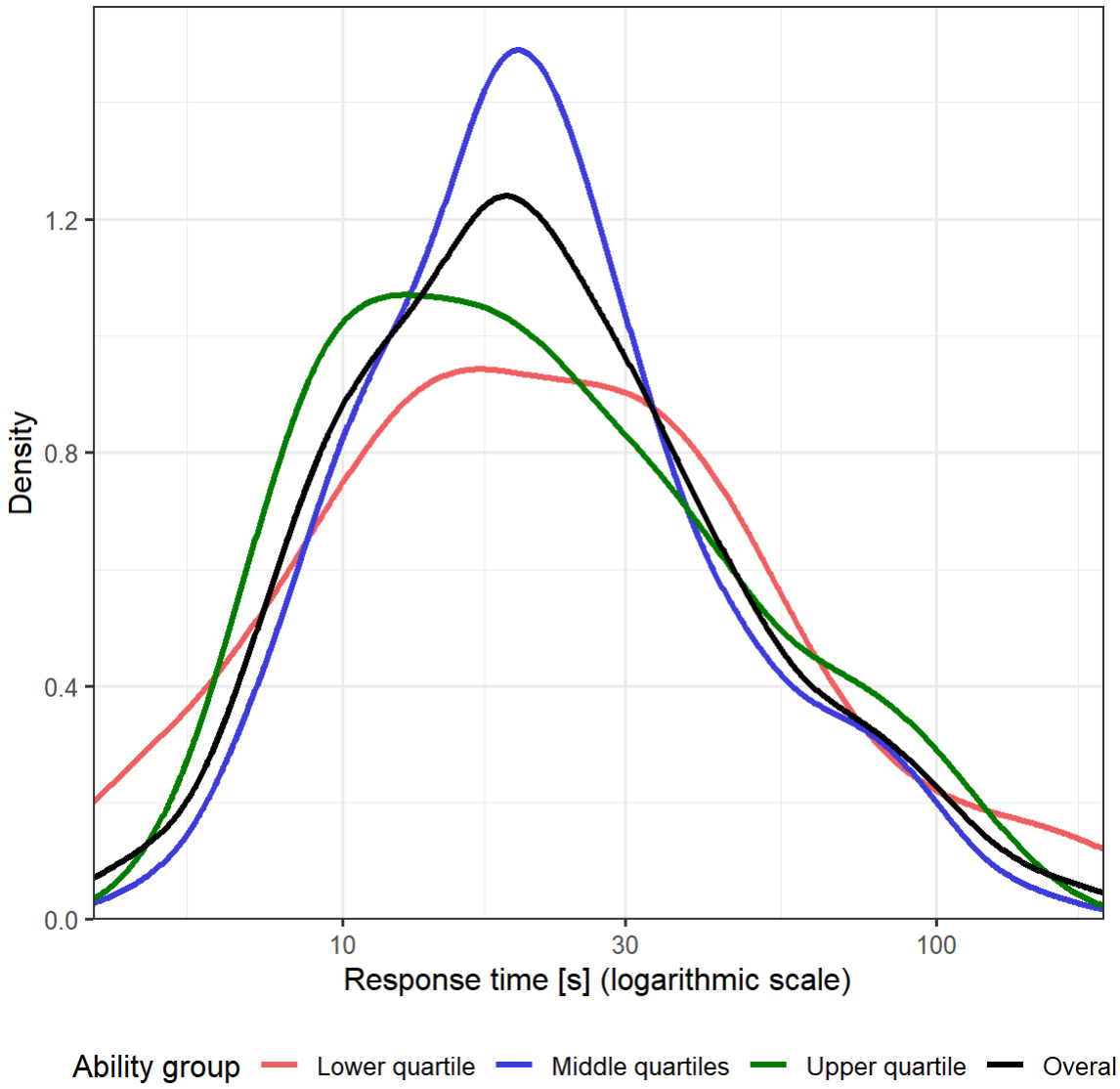
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
48	7	7	0	14	20.0%	28.0%	0.0%	21.5%
54	15	13	5	33	42.9%	52.0%	100.0%	50.8%
64	13	5	0	18	37.1%	20.0%	0.0%	27.7%
Overall	35	25	5	65	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



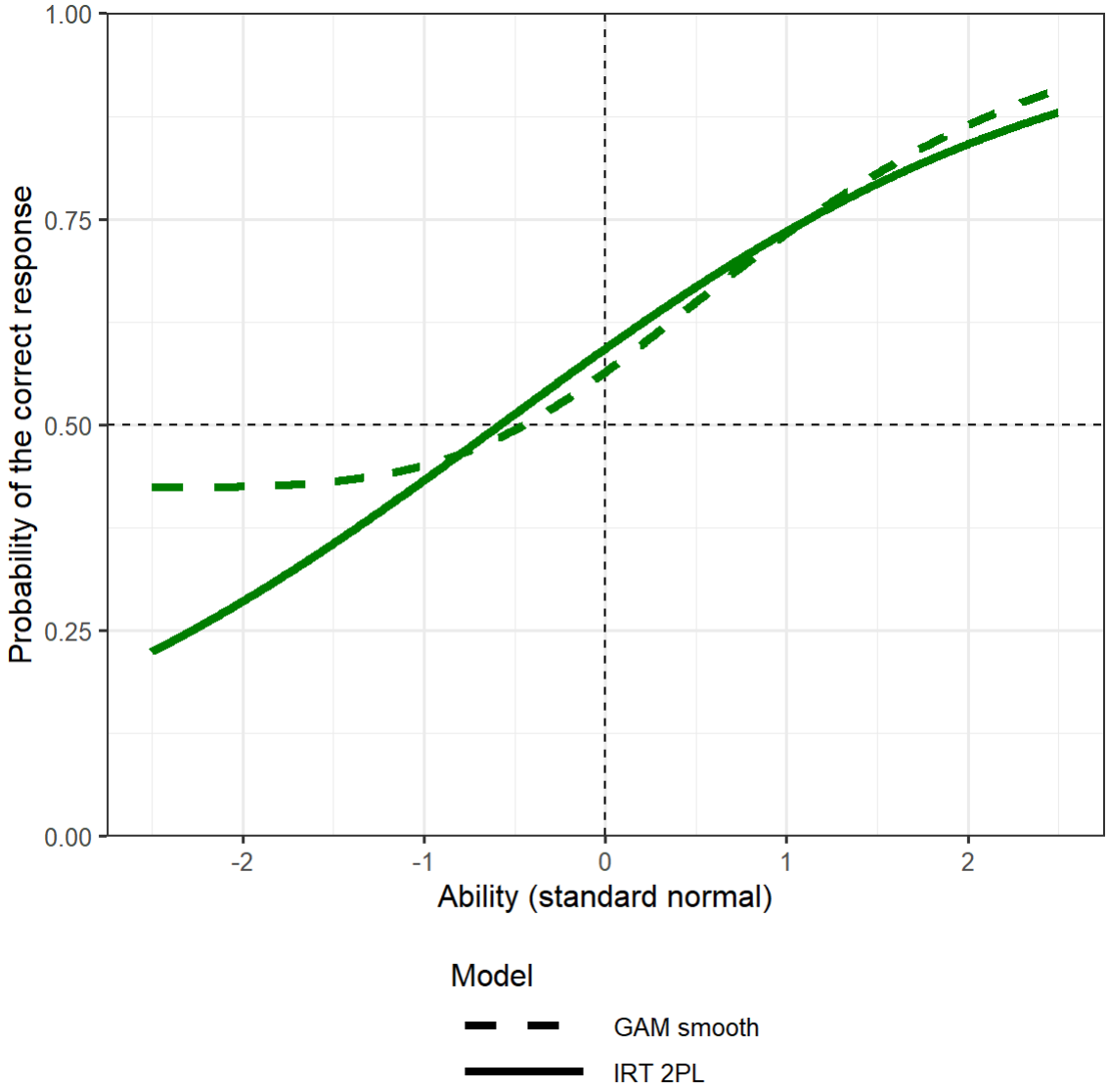
10.1.39.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.8	11.9	20.2	38.4	191.1	3.1	0.9
Middle quartiles	4.5	13.1	20.1	31.2	150.6	3.1	0.7
Upper quartile	5.9	11.0	18.8	36.5	120.5	3.0	0.8
Overall	3.8	12.2	19.8	32.7	191.1	3.1	0.8

10.1.40 M041048 (TIMSS)

10.1.40.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	59.06%	299	TIMSS	MC	0.25	0.39	0.65	−0.58	0.021

10.1.40.2 Gender DIF:

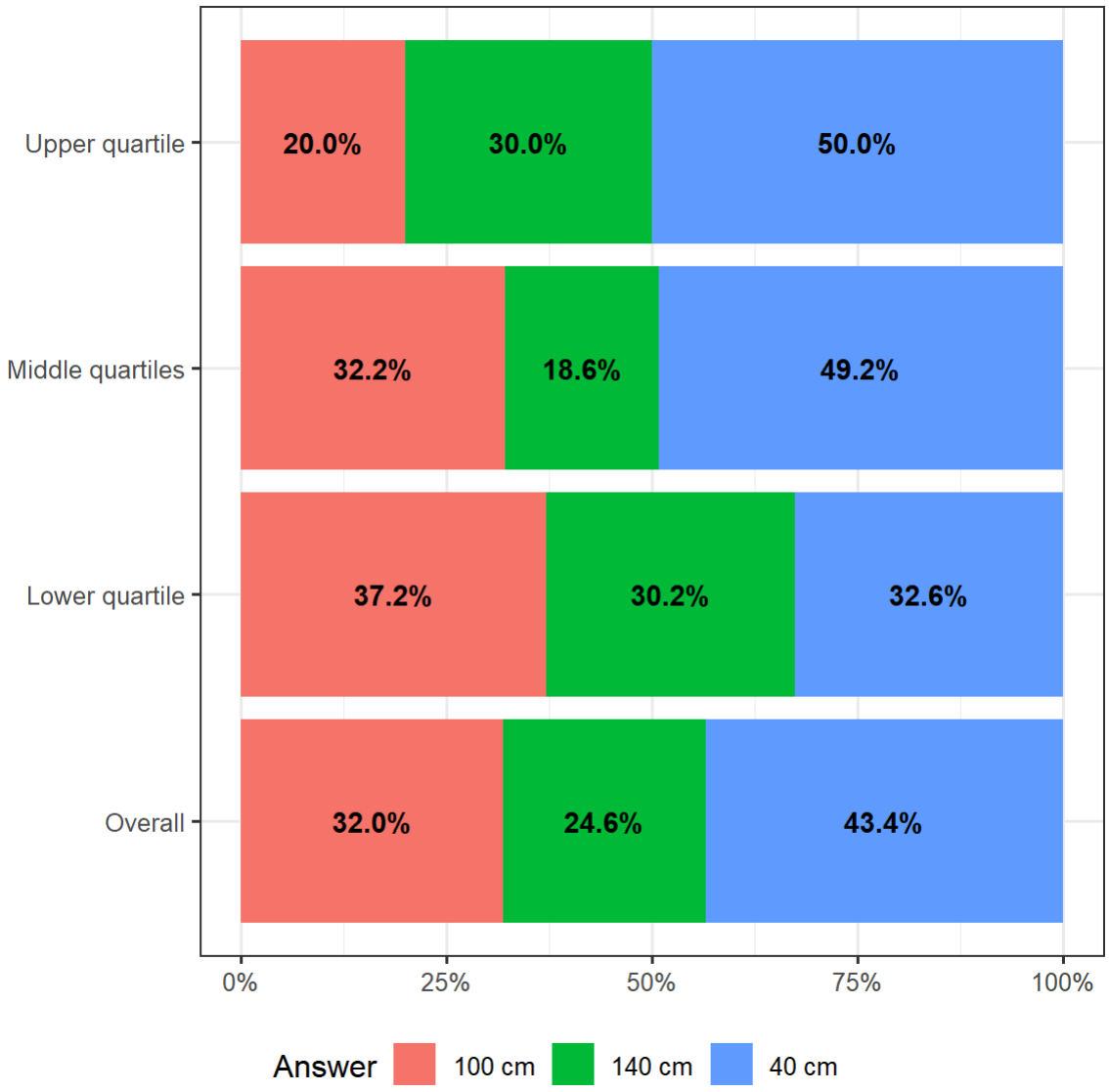
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.82	0.148	1.000	0.011	N/A

10.1.40.3 Frequency of answers:

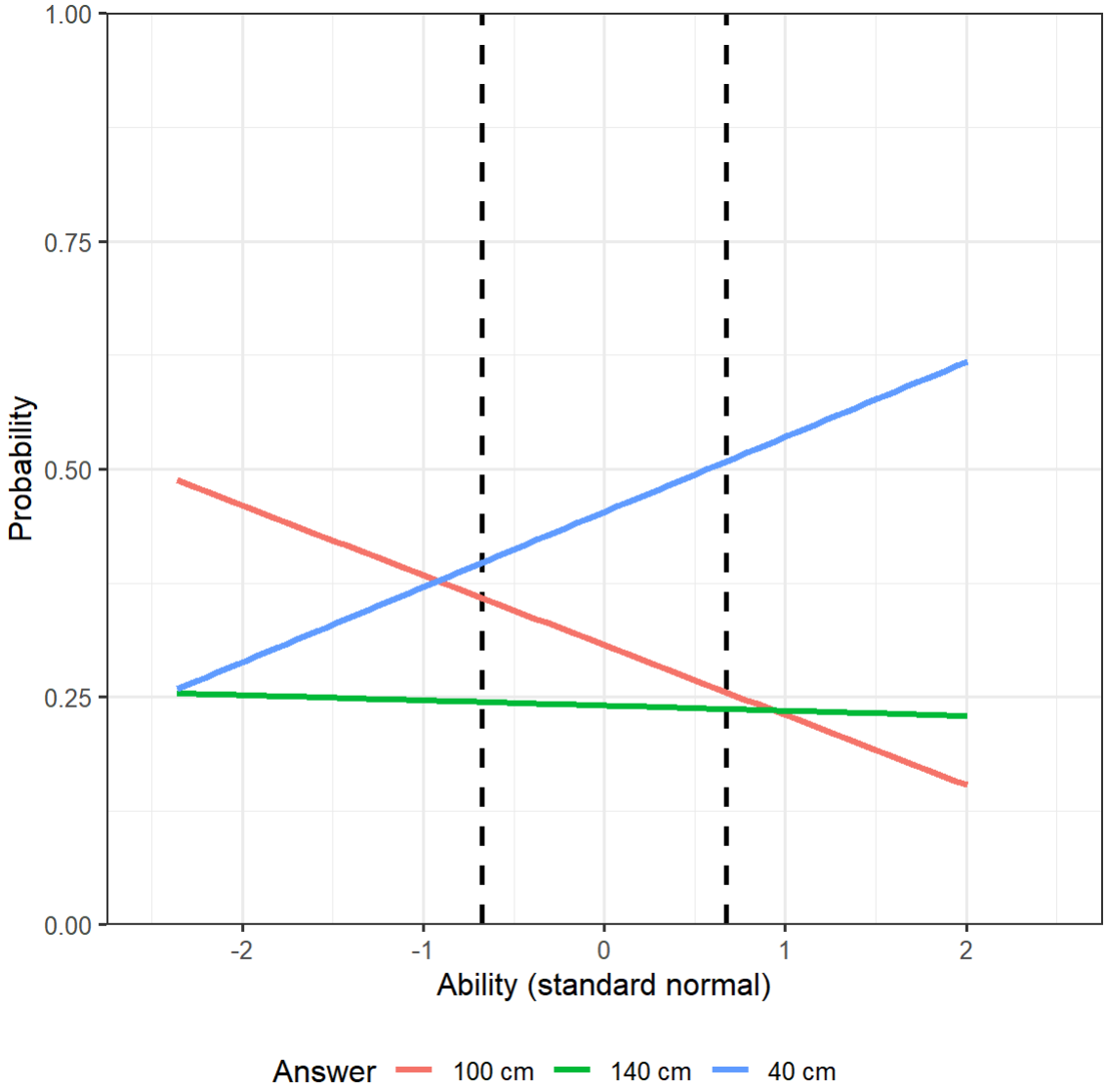
Answer	Frequency	Percent
100 cm	39	13.1%
140 cm	30	10.1%
40 cm	53	17.8%
80 cm (correct)	176	59.1%
Overall	298	100.0%

10.1.40.4 Distractors functioning:

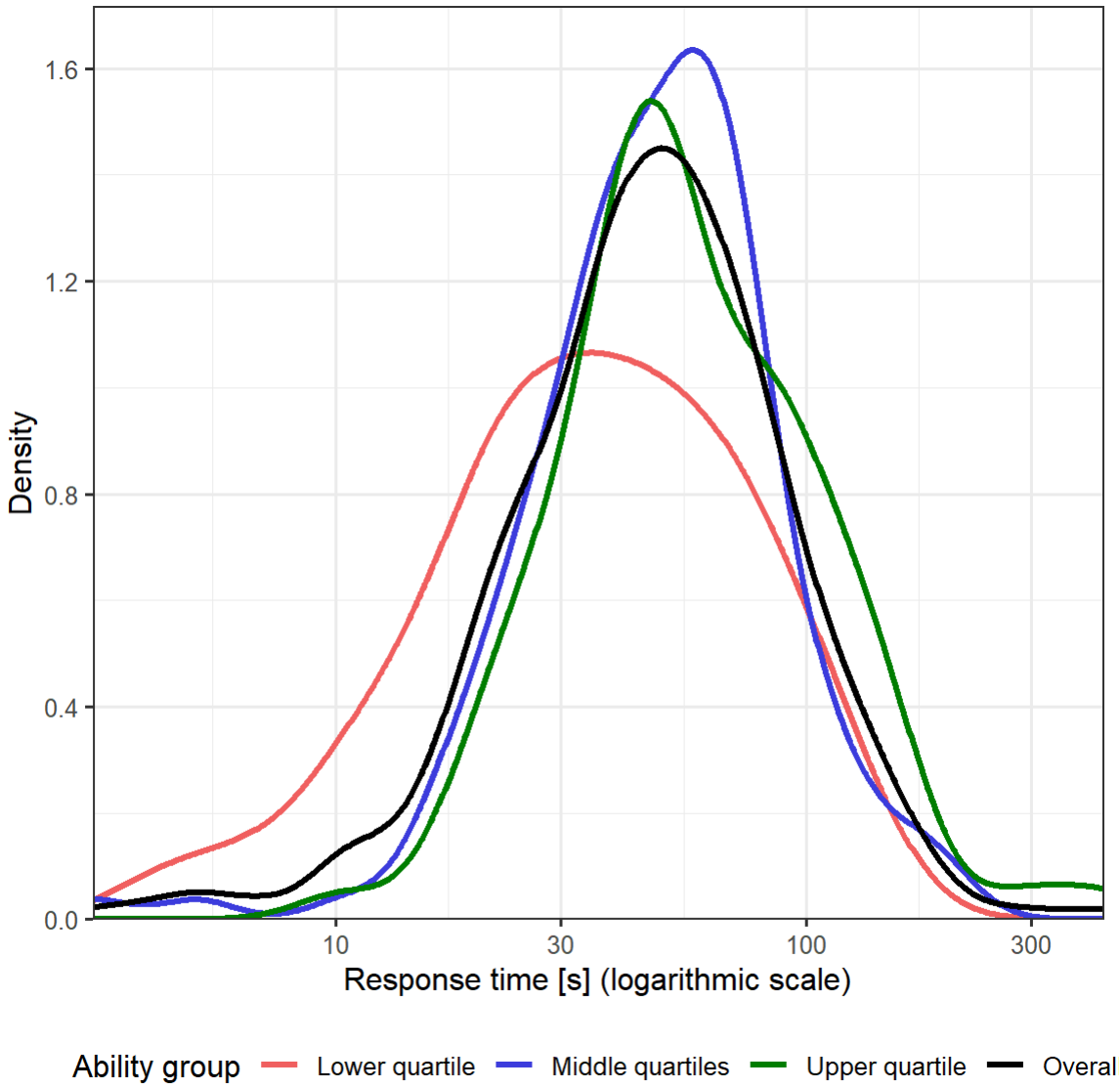
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
100 cm	16	19	4	39	37.2%	32.2%	20.0%	32.0%
140 cm	13	11	6	30	30.2%	18.6%	30.0%	24.6%
40 cm	14	29	10	53	32.6%	49.2%	50.0%	43.4%
Overall	43	59	20	122	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



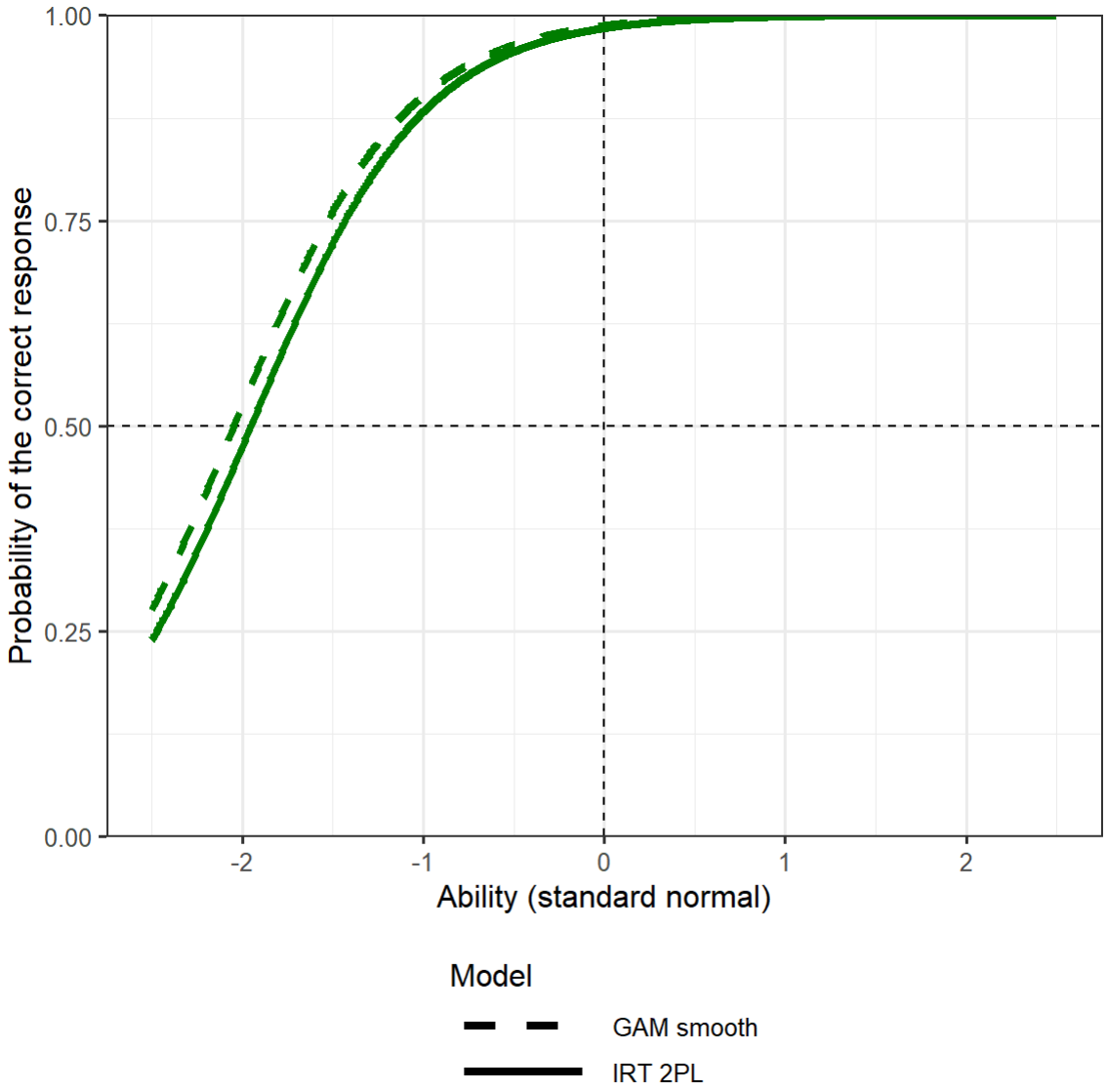
10.1.40.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.0	21.6	36.1	59.0	139.4	3.5	0.8
Middle quartiles	3.0	33.2	49.5	68.8	206.9	3.8	0.6
Upper quartile	10.4	38.0	51.8	85.6	426.4	4.0	0.6
Overall	3.0	30.6	47.6	70.8	426.4	3.8	0.7

10.1.41 M890039 (LLM)

10.1.41.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	93.21%	281	LLM complex setup	MC	0.44	0.23	2.11	−1.95	0.00

10.1.41.2 Gender DIF:

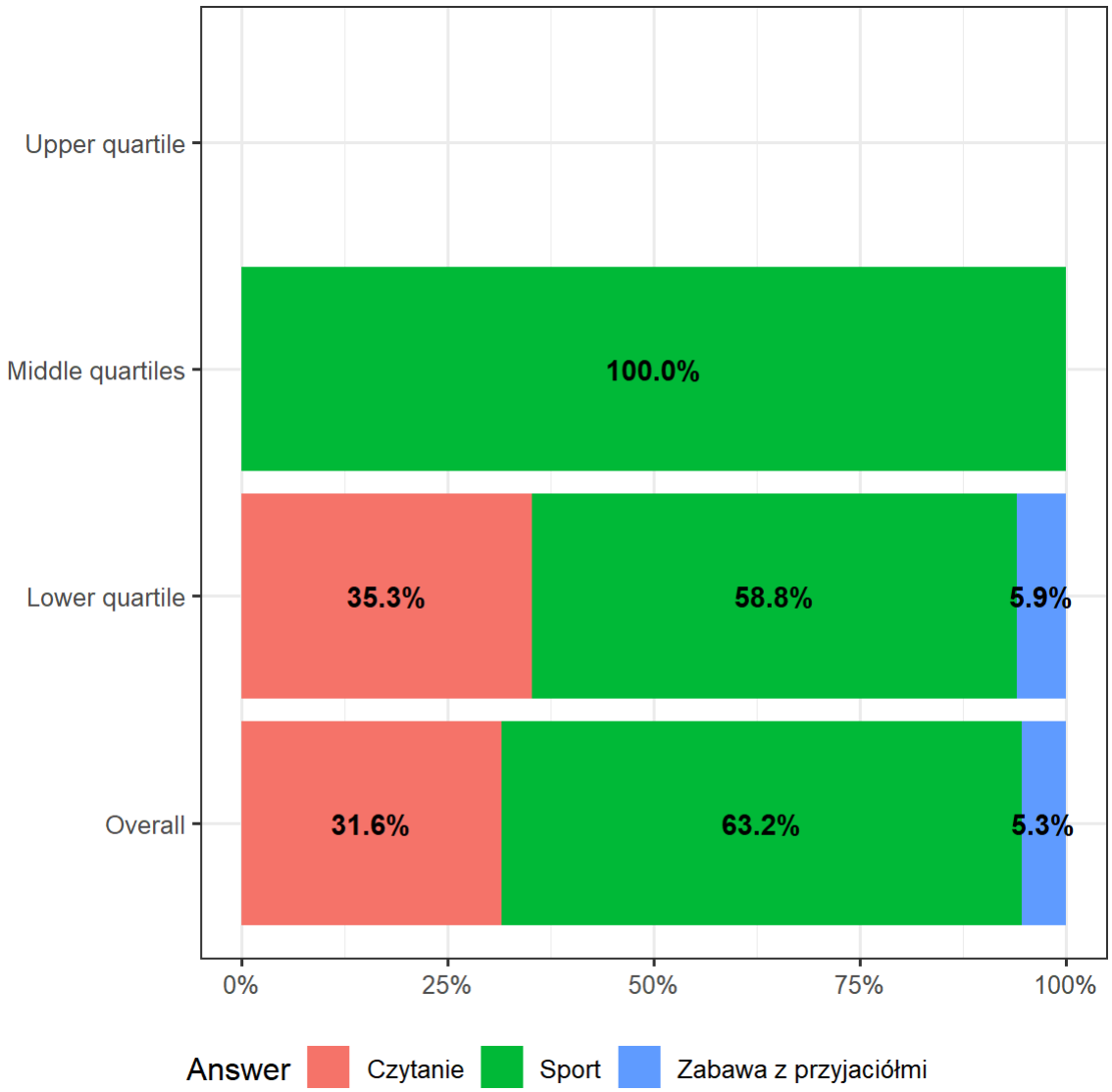
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.97	0.373	1.000	0.014	N/A

10.1.41.3 Frequency of answers:

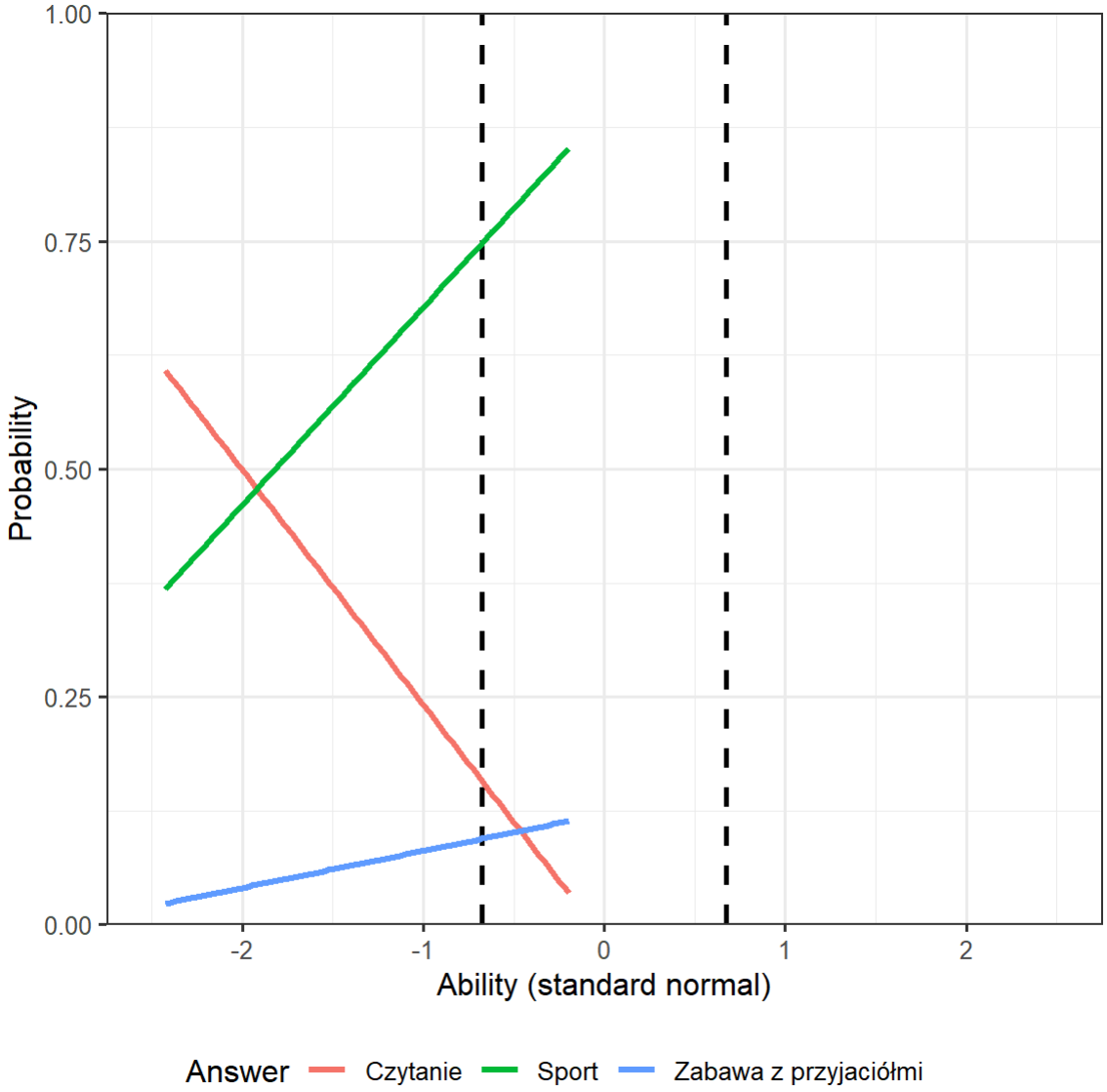
Answer	Frequency	Percent
Czytanie	6	2.1%
Nauka (correct)	261	93.2%
Sport	12	4.3%
Zabawa z przyjaciółmi	1	0.4%
Overall	280	100.0%

10.1.41.4 Distractors functioning:

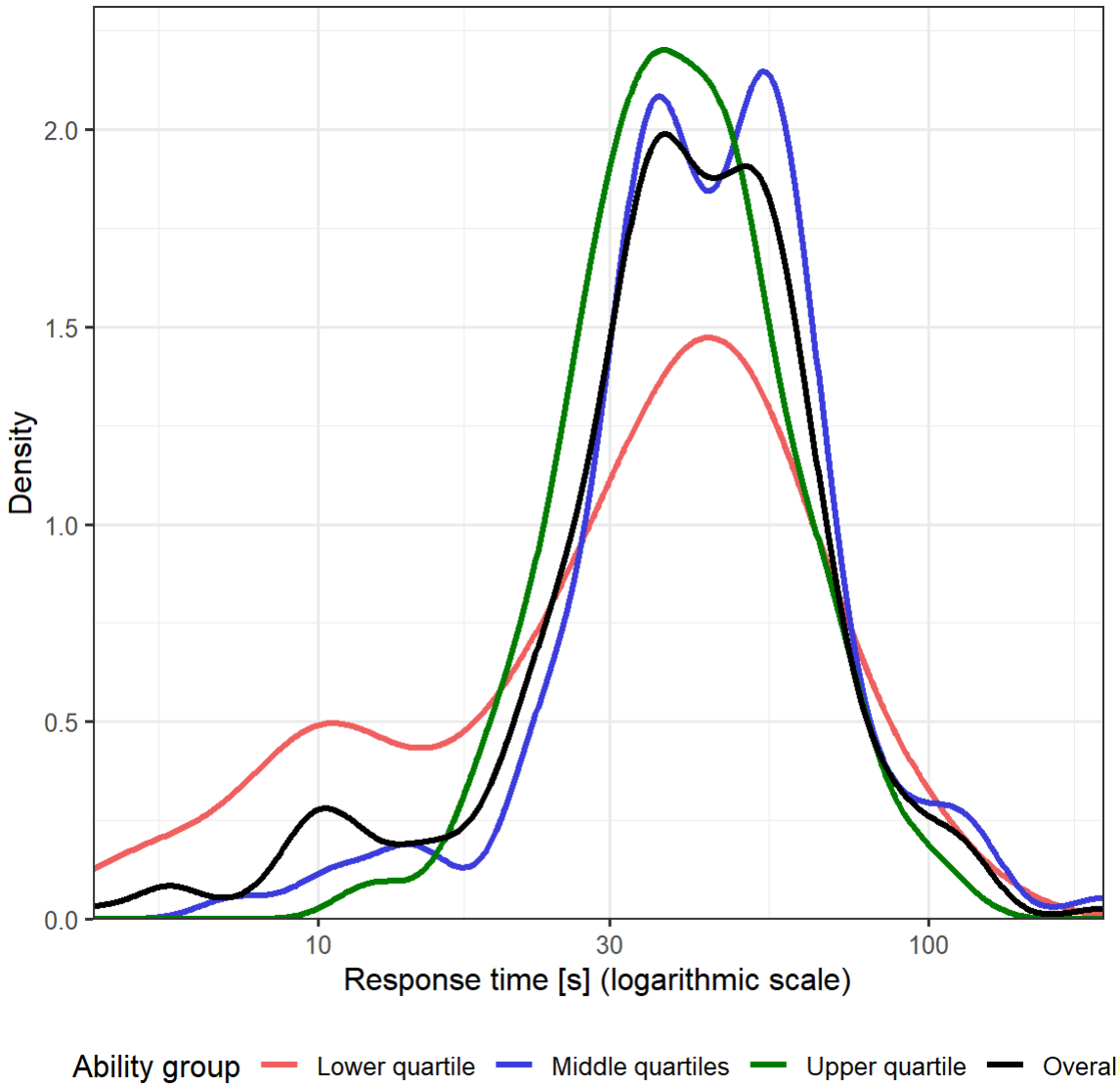
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Czytanie	6	0	0	6	35.3%	0.0%	0.0%	31.6%
Sport	10	2	0	12	58.8%	100.0%	0.0%	63.2%
Zabawa z przyjaciółmi	1	0	0	1	5.9%	0.0%	0.0%	5.3%
Overall	17	2	0	19	100.0%	100.0%	0.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



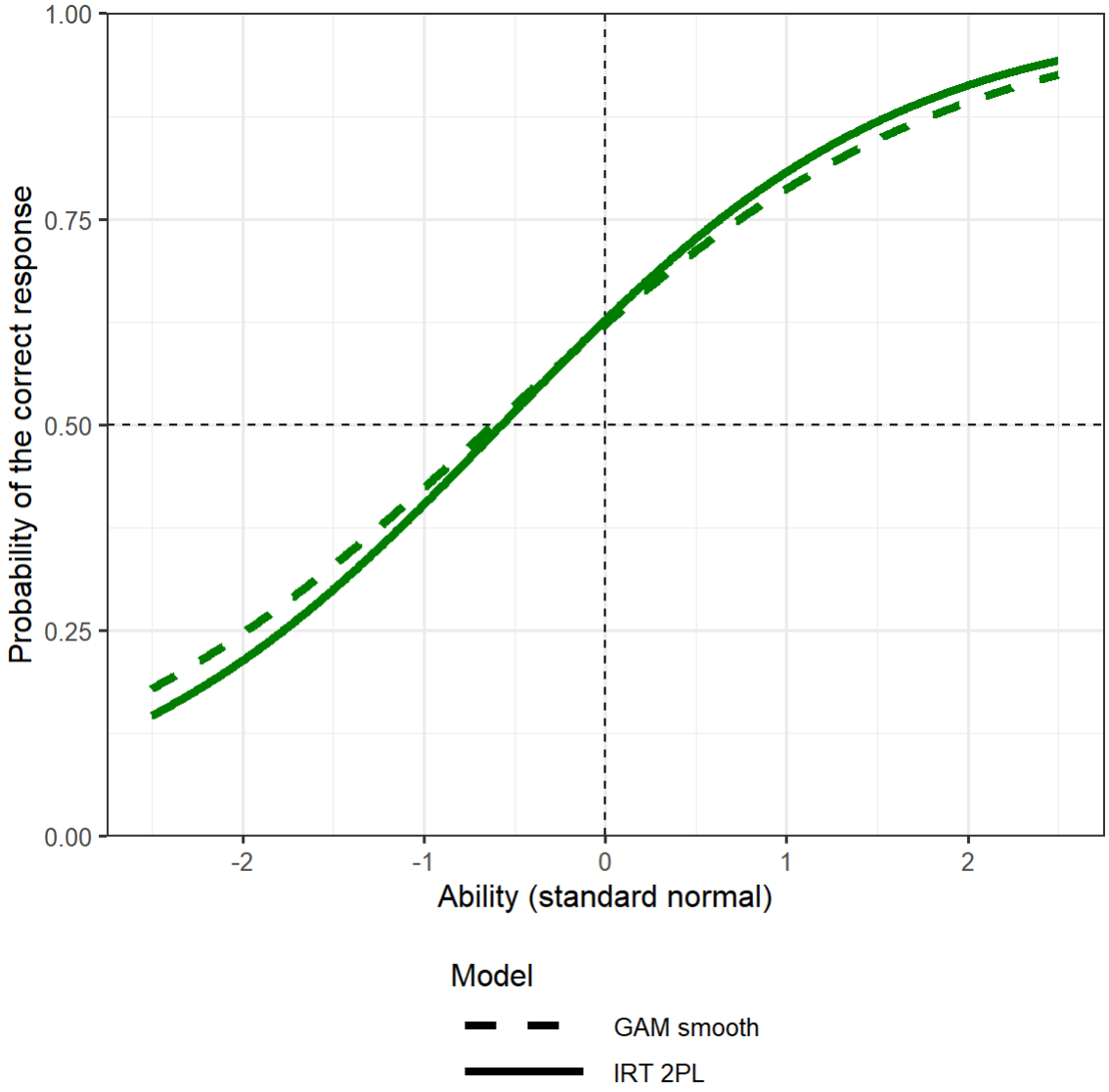
10.1.41.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.3	20.6	37.3	52.8	113.2	3.4	0.8
Middle quartiles	7.3	33.7	43.4	57.0	192.9	3.7	0.5
Upper quartile	12.3	30.3	37.8	49.3	105.0	3.7	0.4
Overall	4.3	30.6	39.6	53.8	192.9	3.6	0.6

10.1.42 M041169 (TIMSS)

10.1.42.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	60.00%	306	TIMSS	MC	0.35	0.38	0.91	−0.57	0.016

10.1.42.2 Gender DIF:

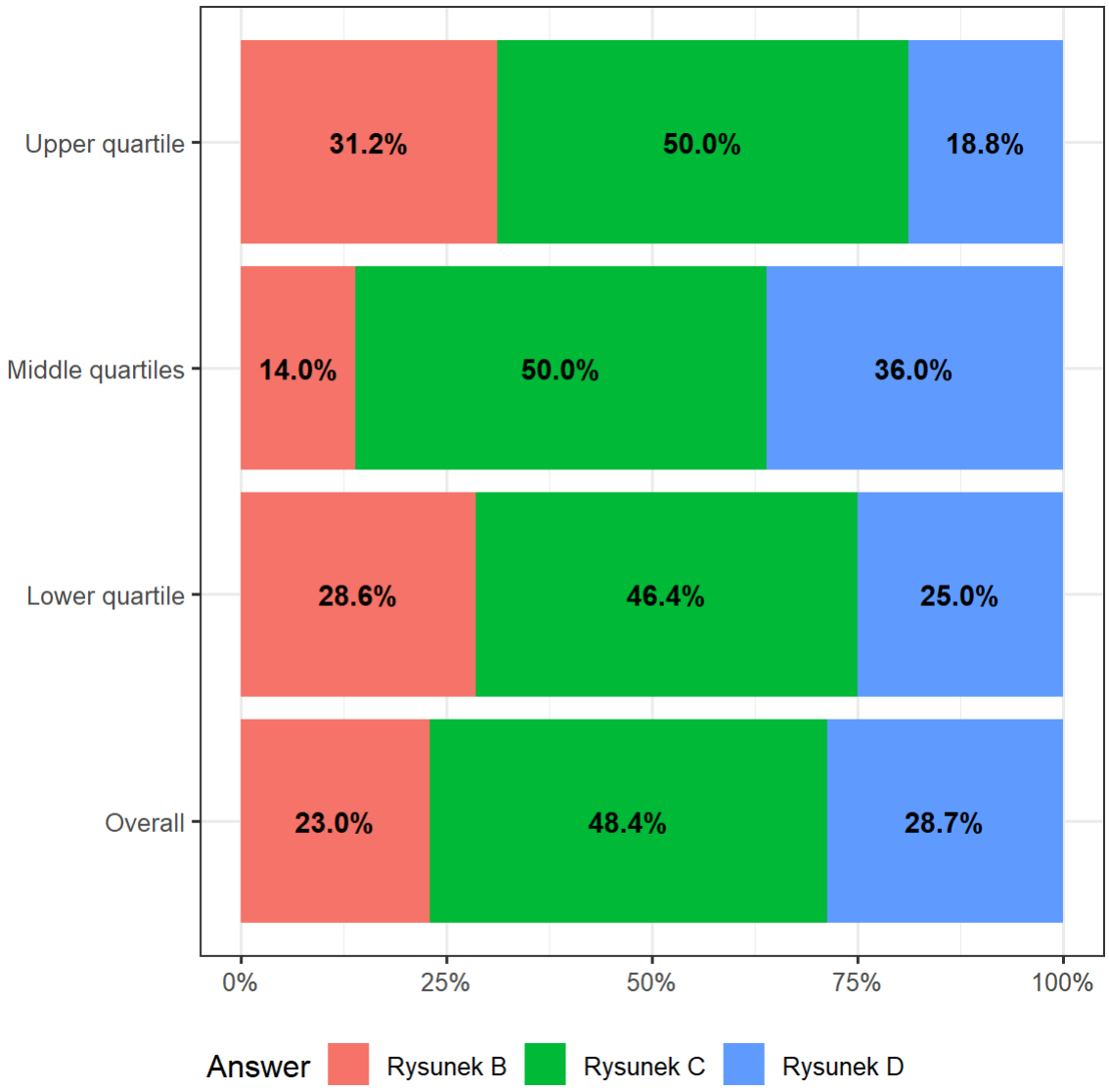
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.50	0.473	1.000	0.004	N/A

10.1.42.3 Frequency of answers:

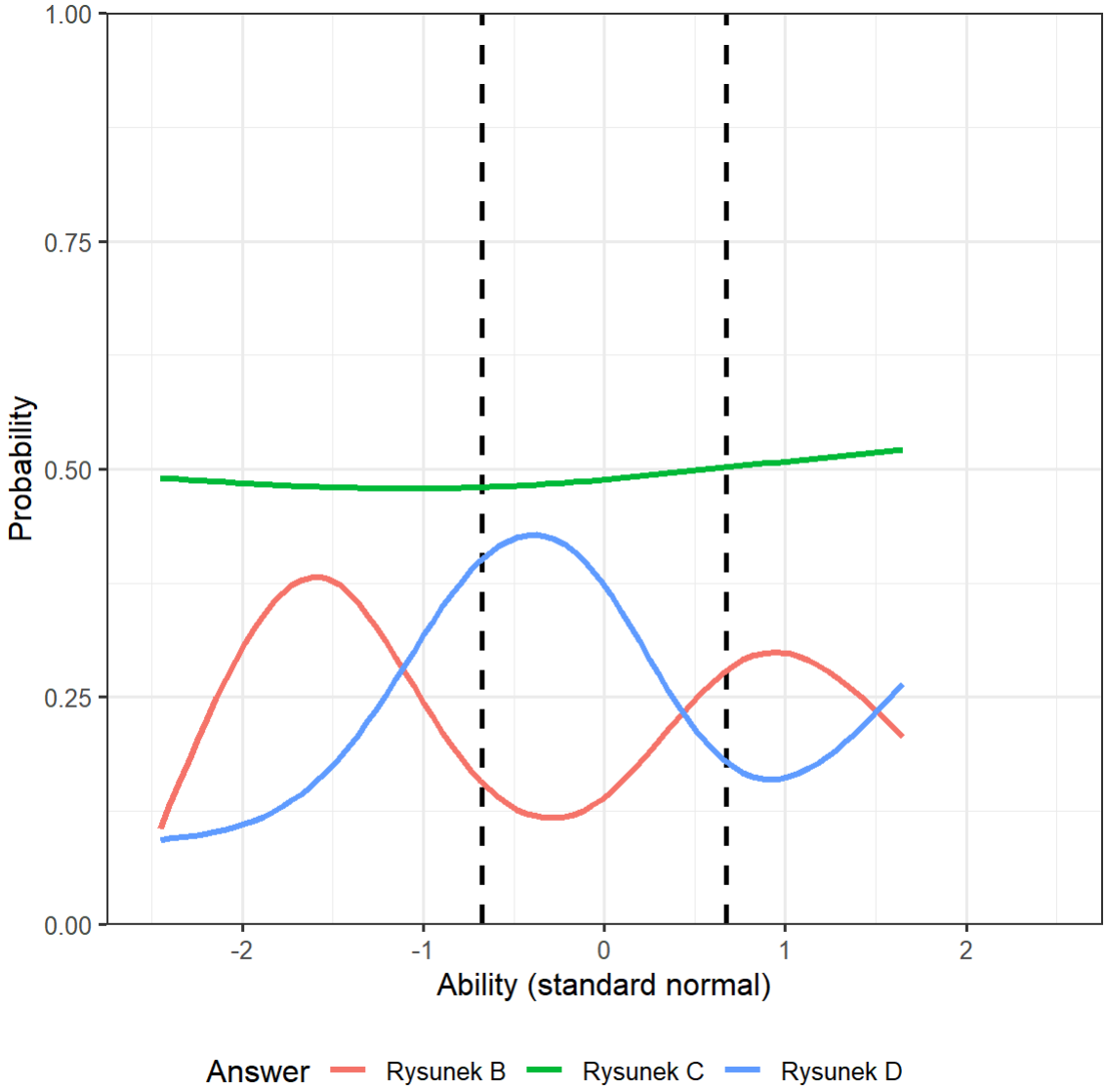
Answer	Frequency	Percent
Rysunek A (correct)	183	60.0%
Rysunek B	28	9.2%
Rysunek C	59	19.3%
Rysunek D	35	11.5%
Overall	305	100.0%

10.1.42.4 Distractors functioning:

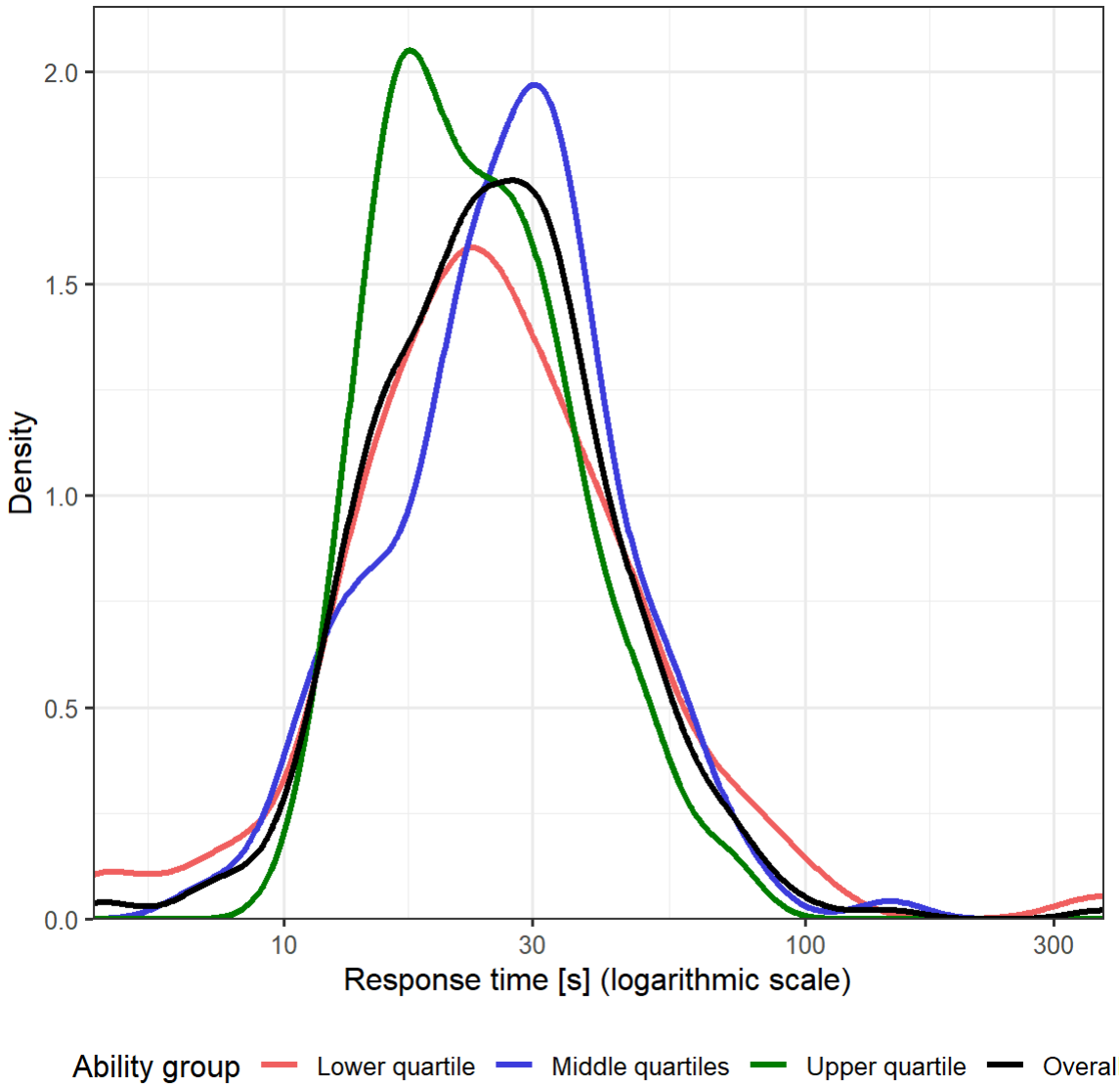
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Rysunek B	16	7	5	28	28.6%	14.0%	31.2%	23.0%
Rysunek C	26	25	8	59	46.4%	50.0%	50.0%	48.4%
Rysunek D	14	18	3	35	25.0%	36.0%	18.8%	28.7%
Overall	56	50	16	122	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



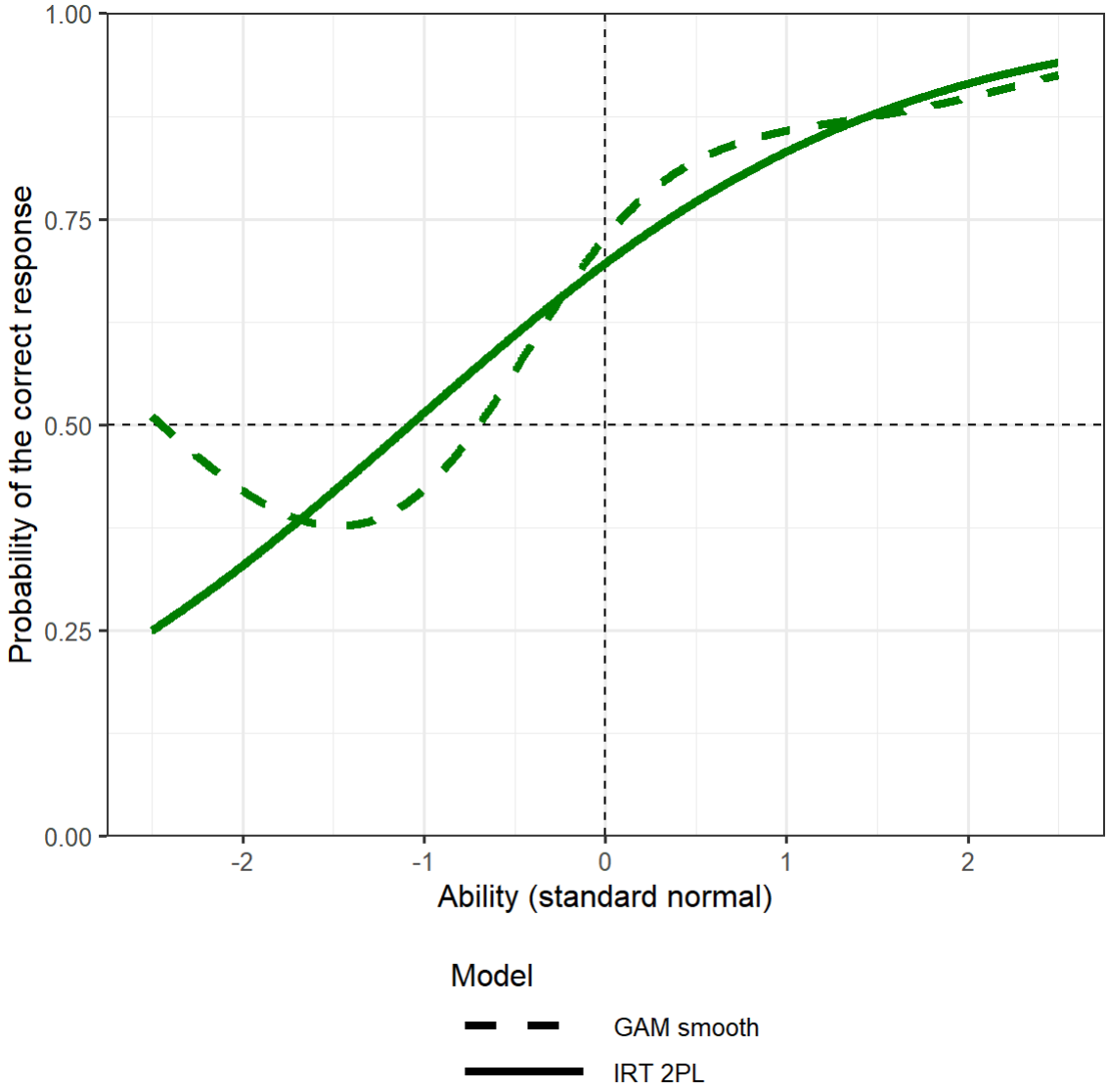
10.1.42.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.3	16.9	23.8	35.5	373.4	3.2	0.7
Middle quartiles	6.7	20.0	28.0	36.8	145.6	3.3	0.5
Upper quartile	11.2	16.7	22.5	30.6	69.6	3.2	0.4
Overall	4.3	17.7	24.7	34.7	373.4	3.2	0.5

10.1.43 M890025 (LLM)

10.1.43.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	67.71%	319	LLM complex setup	MC	0.35	0.22	0.77	−1.07	0.00

10.1.43.2 Gender DIF:

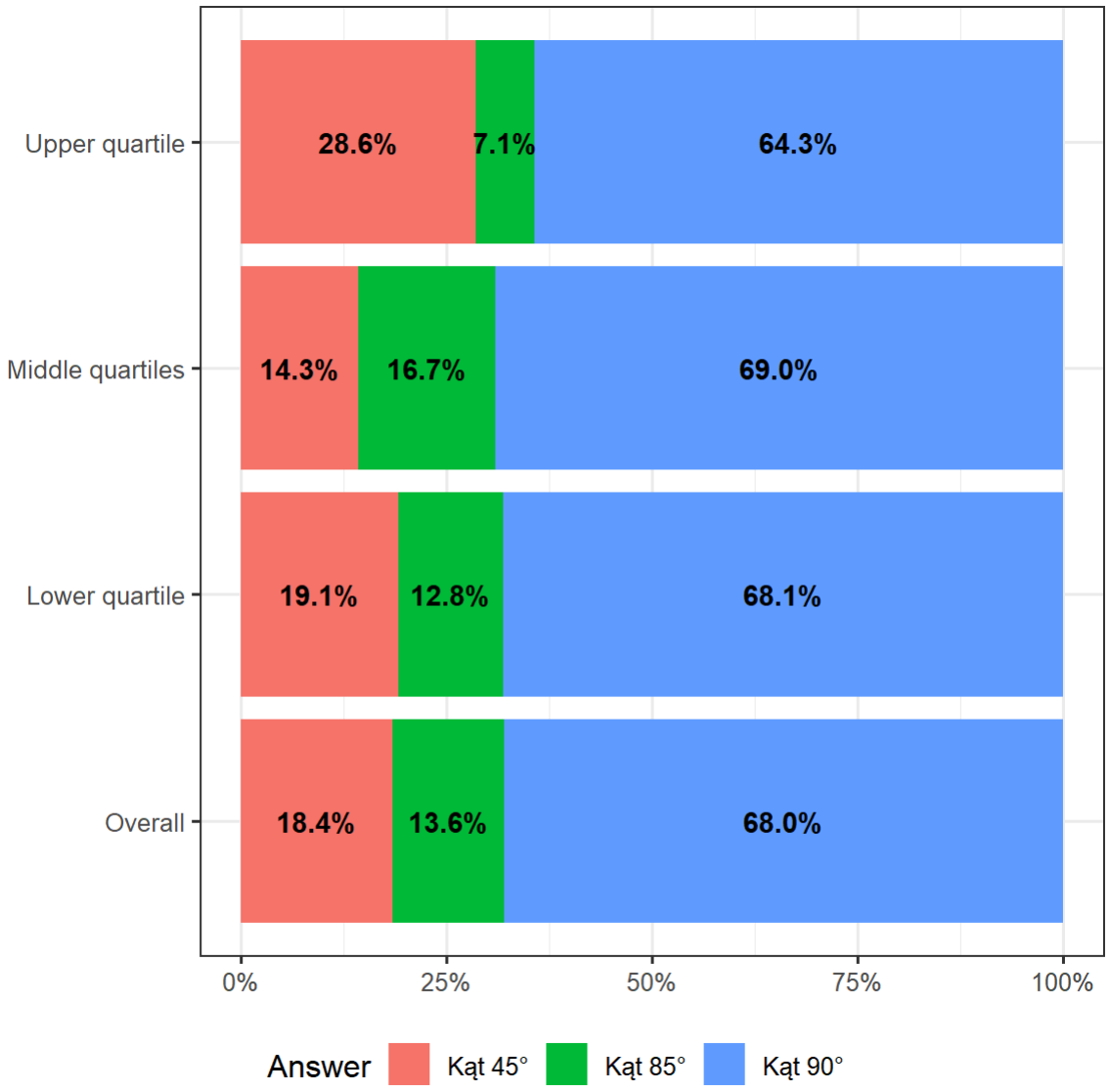
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.83	0.400	1.000	0.005	N/A

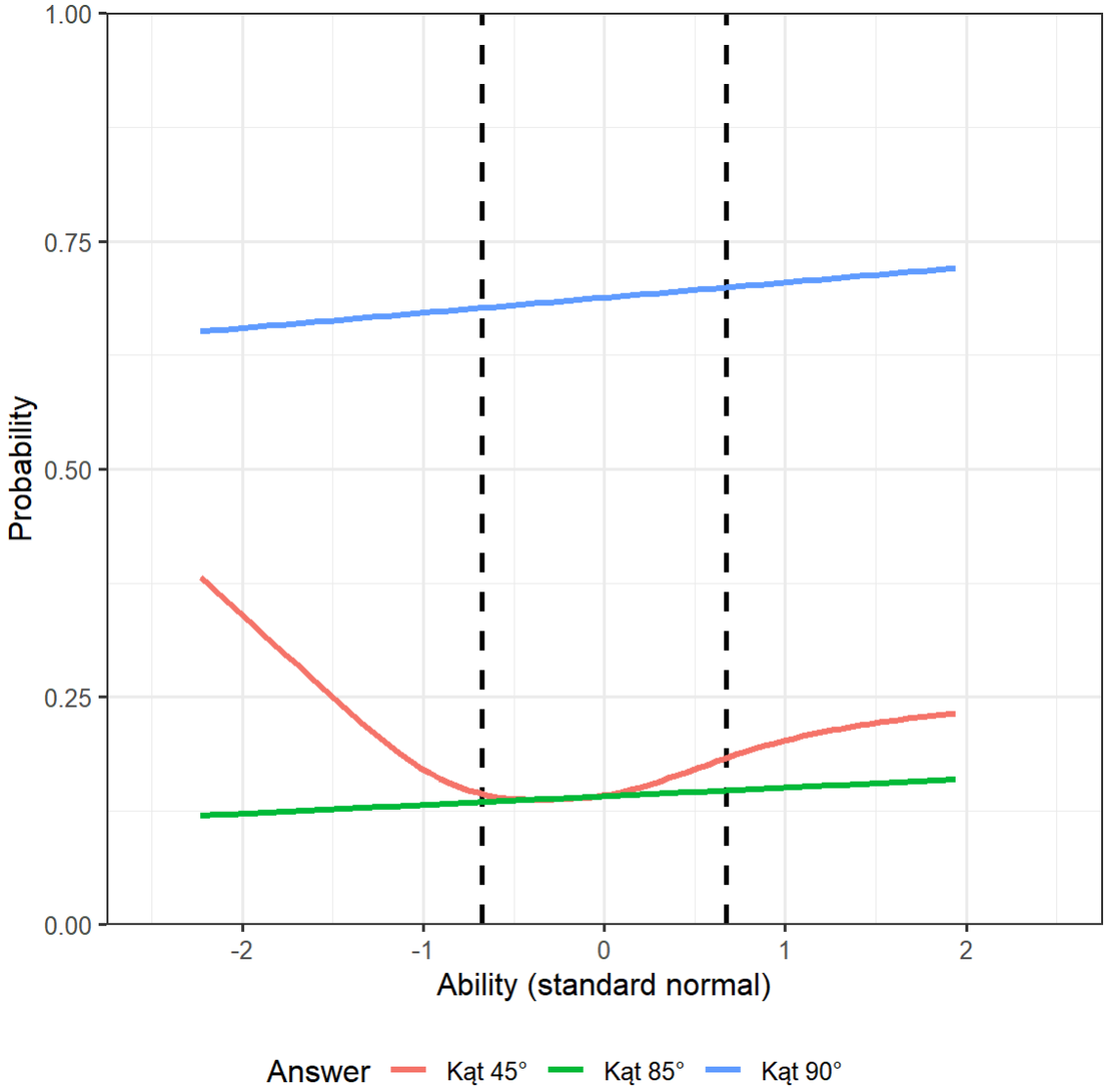
10.1.43.3 Frequency of answers:

Answer	Frequency	Percent
Kąt 120° (correct)	216	67.7%
Kąt 45°	19	6.0%
Kąt 85°	14	4.4%
Kąt 90°	70	21.9%
Overall	319	100.0%

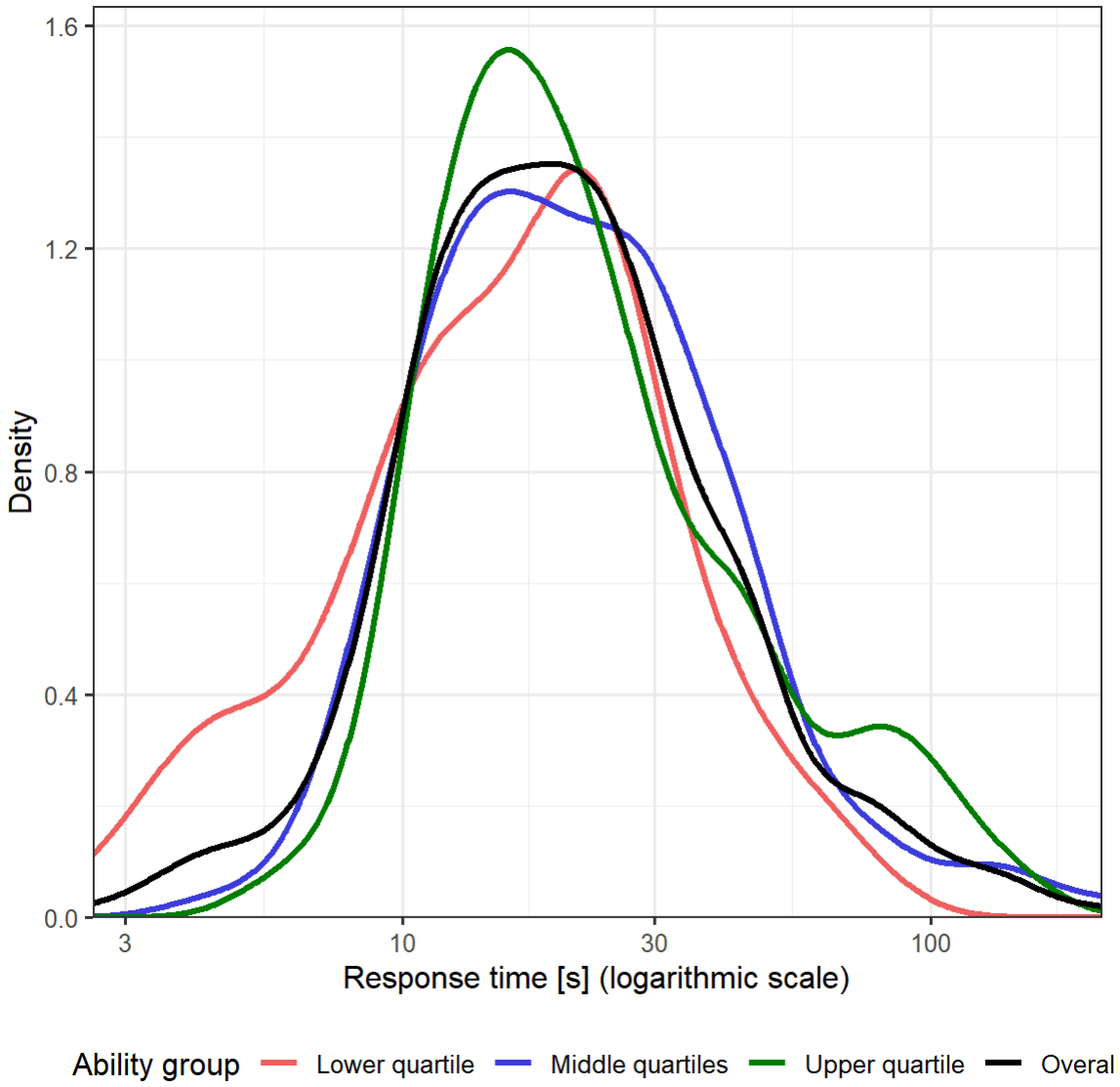
10.1.43.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Kąt 45°	9	6	4	19	19.1%	14.3%	28.6%	18.4%
Kąt 85°	6	7	1	14	12.8%	16.7%	7.1%	13.6%
Kąt 90°	32	29	9	70	68.1%	69.0%	64.3%	68.0%
Overall	47	42	14	103	100.0%	100.0%	100.0%	100.0%


Upper quartile
Middle quartiles
Lower quartile
Overall
28.6%
7.1%
64.3%
14.3%
16.7%
69.0%
19.1%
12.8%
68.1%
18.4%
13.6%
68.0%
0%
25%
50%
75%
100%
Answer
Kąt 45°
Kąt 85°
Kąt 90°



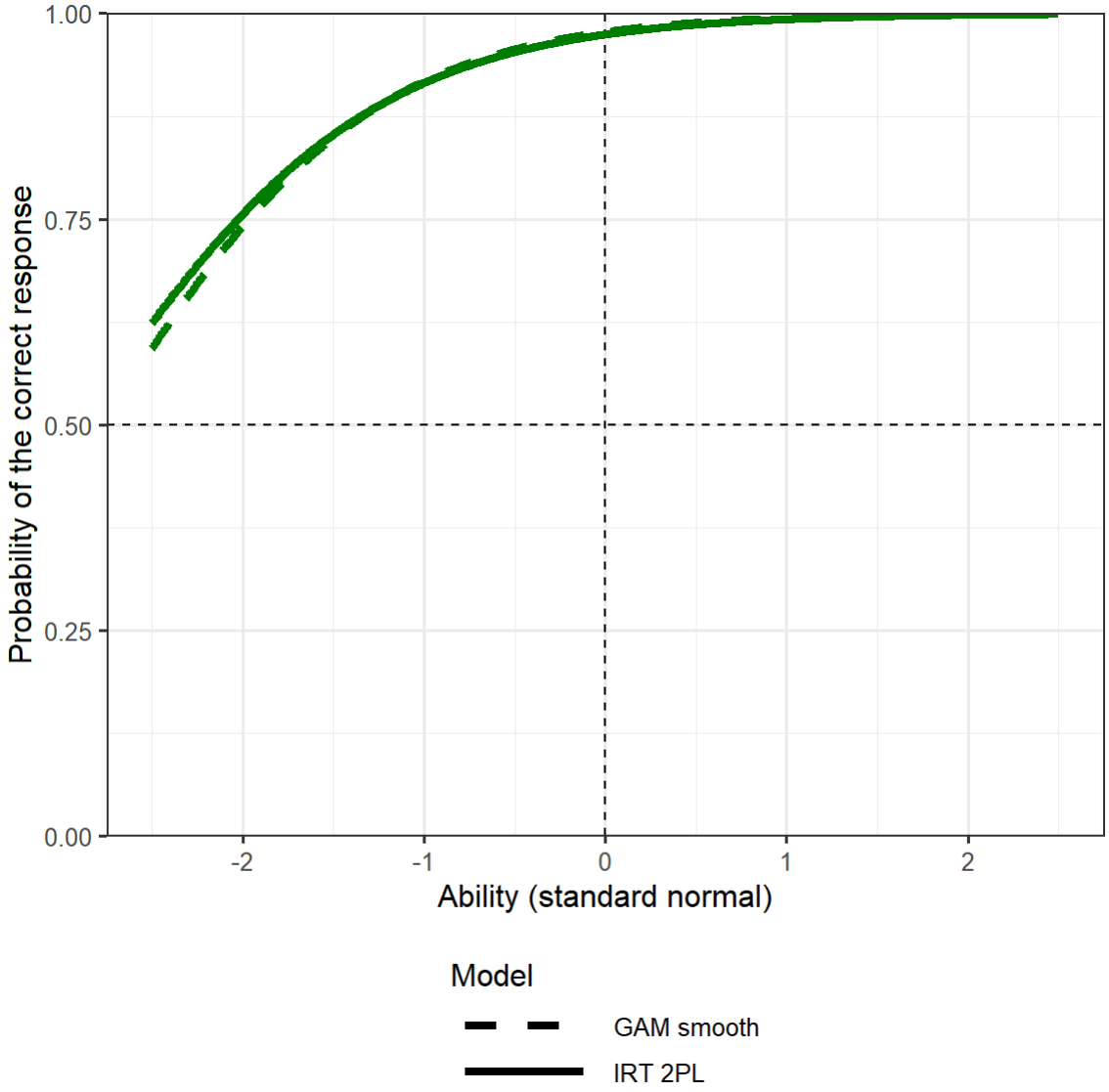
10.1.43.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.6	10.0	17.0	24.1	71.8	2.8	0.7
Middle quartiles	4.3	13.2	20.4	30.6	210.2	3.1	0.7
Upper quartile	5.9	14.1	18.8	31.1	145.5	3.1	0.7
Overall	2.6	12.5	19.7	30.1	210.2	3.0	0.7

10.1.44 M100020 (LLM)

10.1.44.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	94.97%	299	LLM simple setup	MC	0.27	0.20	1.25	−2.90	0.016

10.1.44.2 Gender DIF:

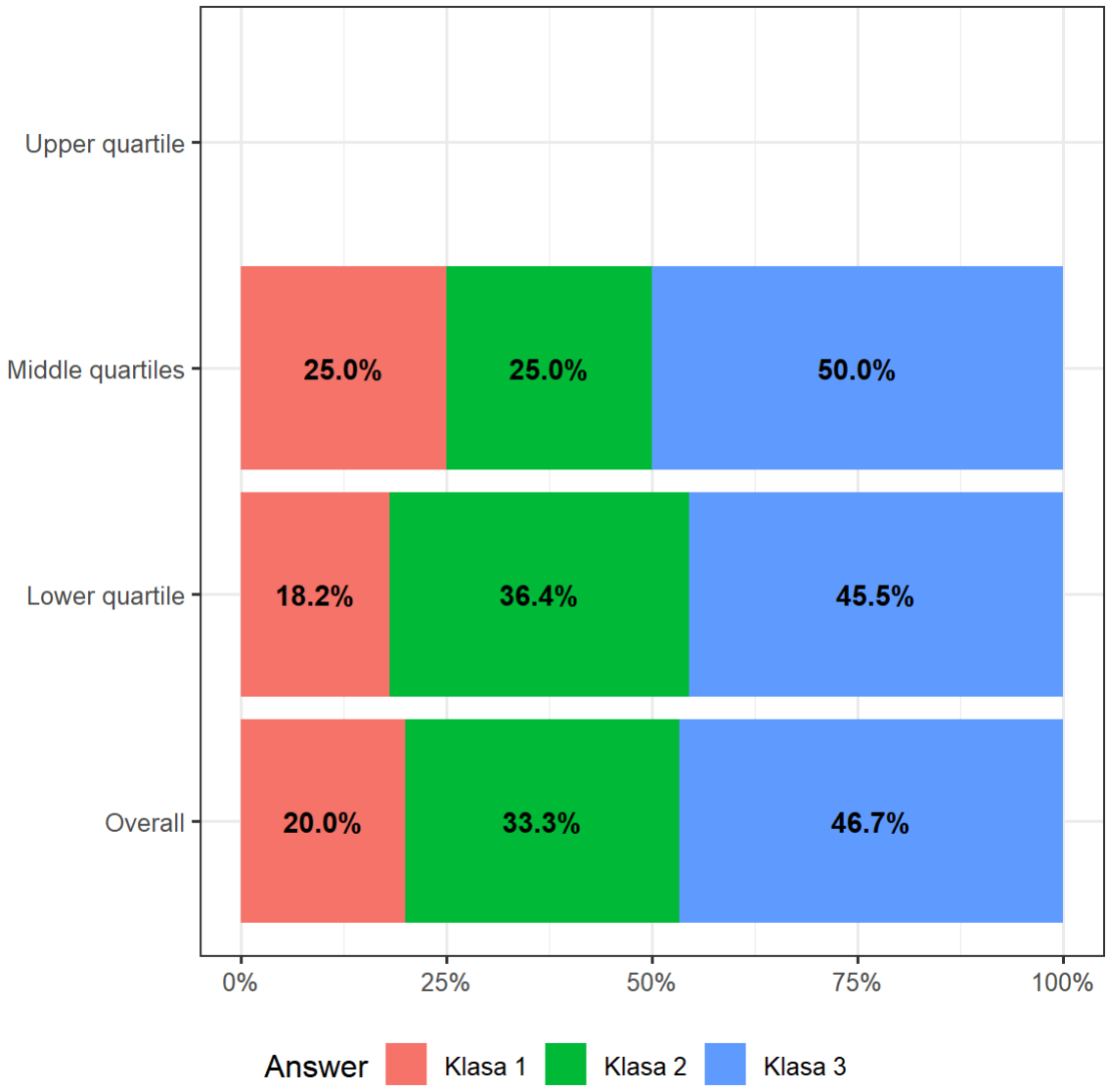
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.05	0.217	1.000	0.026	N/A

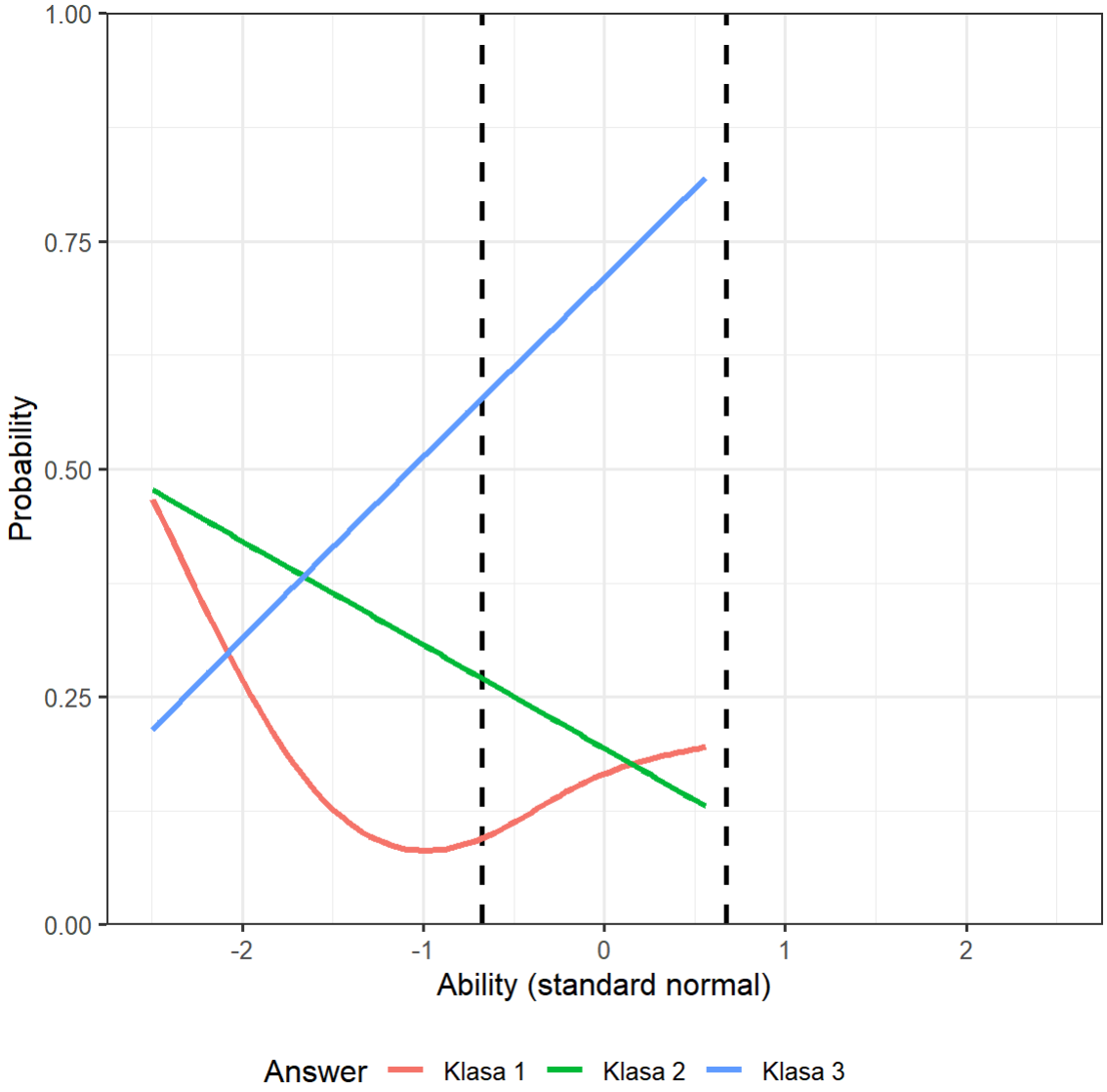
10.1.44.3 Frequency of answers:

Answer	Frequency	Percent
Klasa 1	3	1.0%
Klasa 2	5	1.7%
Klasa 3	7	2.3%
Klasa 4 (correct)	283	95.0%
Overall	298	100.0%

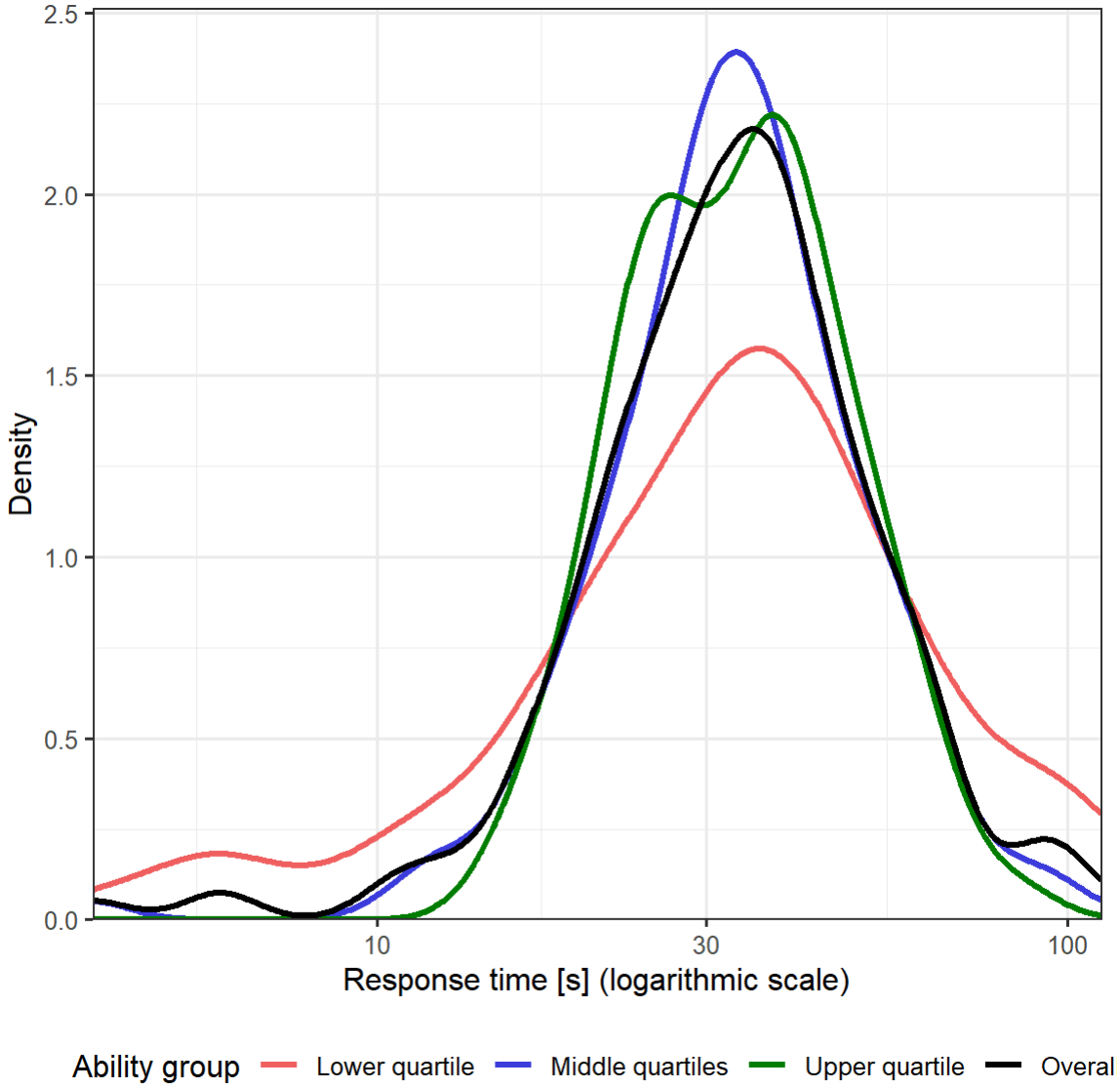
10.1.44.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Klasa 1	2	1	0	3	18.2%	25.0%	0.0%	20.0%
Klasa 2	4	1	0	5	36.4%	25.0%	0.0%	33.3%
Klasa 3	5	2	0	7	45.5%	50.0%	0.0%	46.7%
Overall	11	4	0	15	100.0%	100.0%	0.0%	100.0%


Upper quartile
Middle quartiles
Lower quartile
Overall
25.0%
25.0%
50.0%
18.2%
36.4%
45.5%
20.0%
33.3%
46.7%
0%
25%
50%
75%
100%
Answer
Klasa 1
Klasa 2
Klasa 3



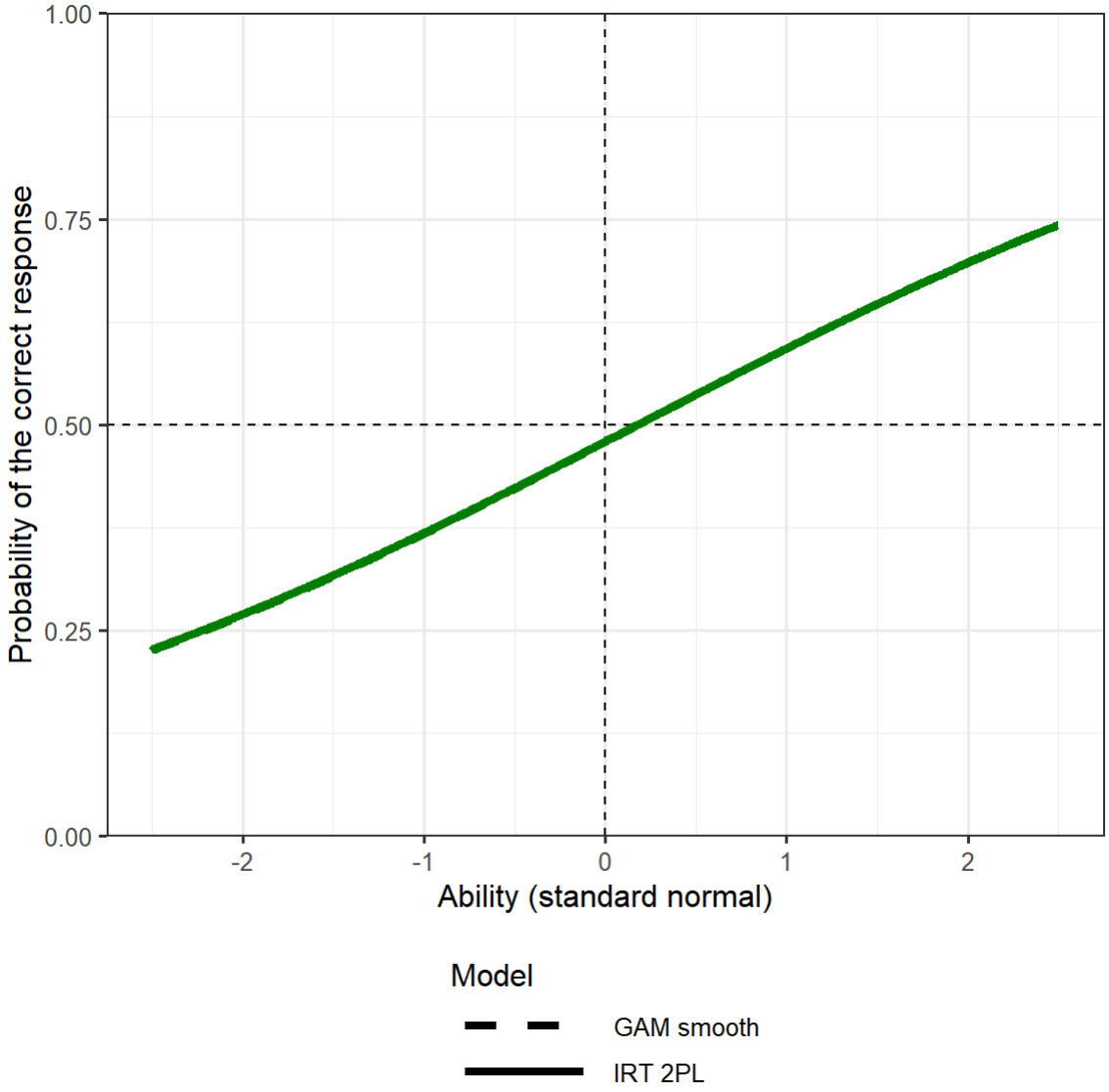
10.1.44.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.9	21.7	34.0	47.6	112.0	3.4	0.7
Middle quartiles	3.9	25.6	33.2	42.7	101.9	3.5	0.4
Upper quartile	15.4	25.1	33.8	42.3	84.8	3.5	0.4
Overall	3.9	24.8	33.5	43.4	112.0	3.5	0.5

10.1.45 M041151 (TIMSS)

10.1.45.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	47.54%	306	TIMSS	MC	0.22	0.15	0.46	0.18	0.010

10.1.45.2 Gender DIF:

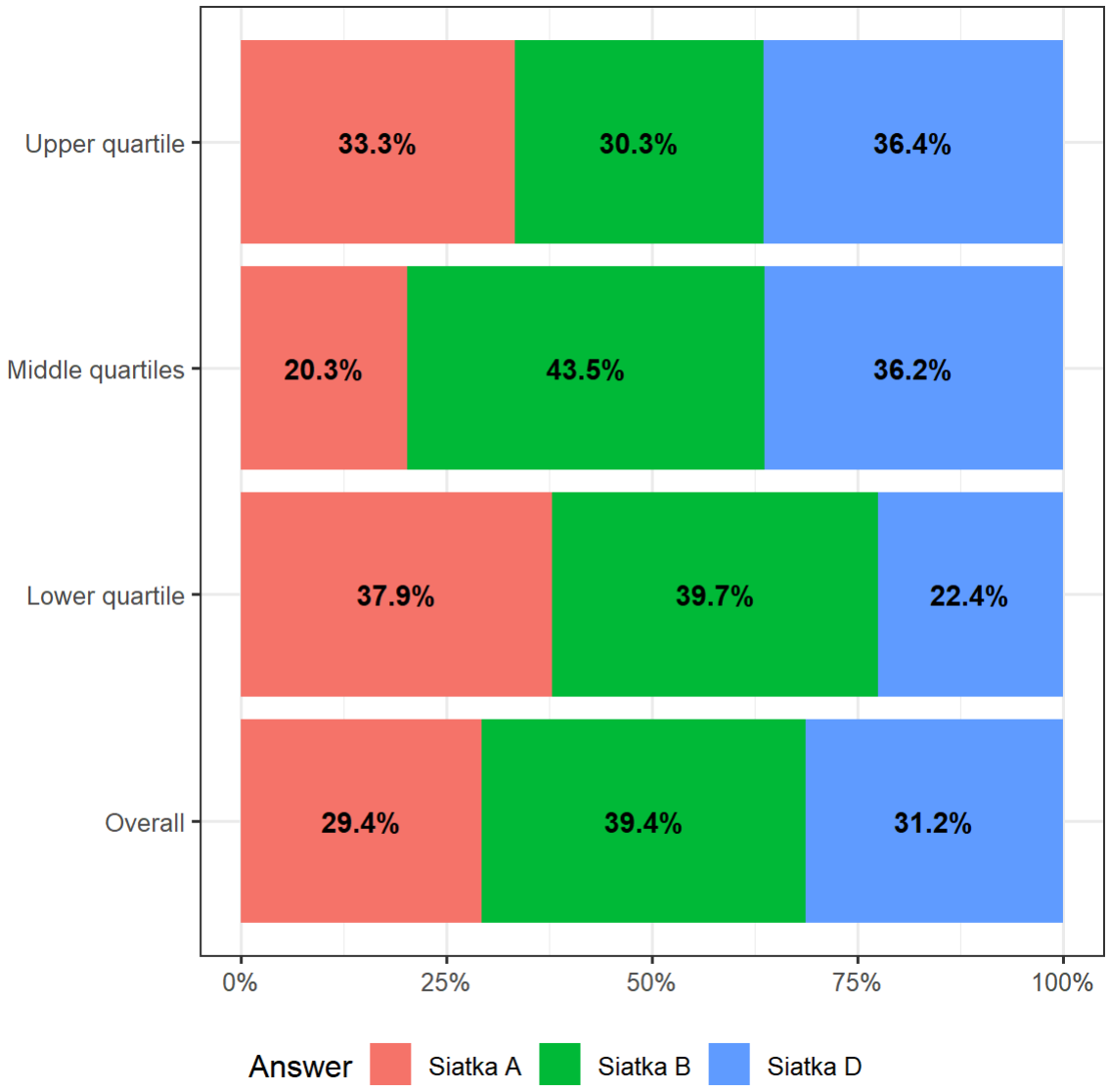
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.25	0.536	1.000	0.003	N/A

10.1.45.3 Frequency of answers:

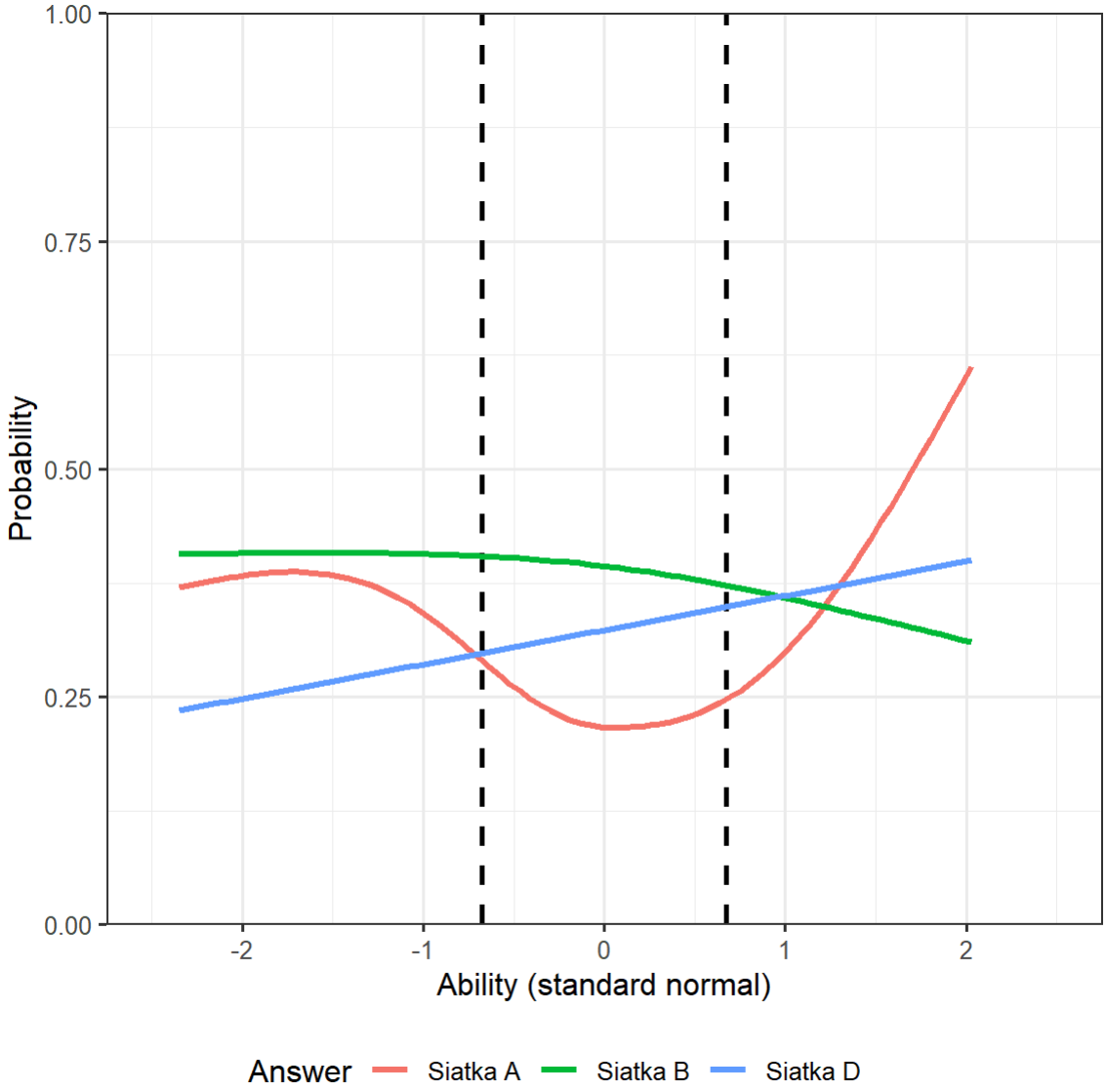
Answer	Frequency	Percent
Siatka A	47	15.4%
Siatka B	63	20.7%
Siatka C (correct)	145	47.5%
Siatka D	50	16.4%
Overall	305	100.0%

10.1.45.4 Distractors functioning:

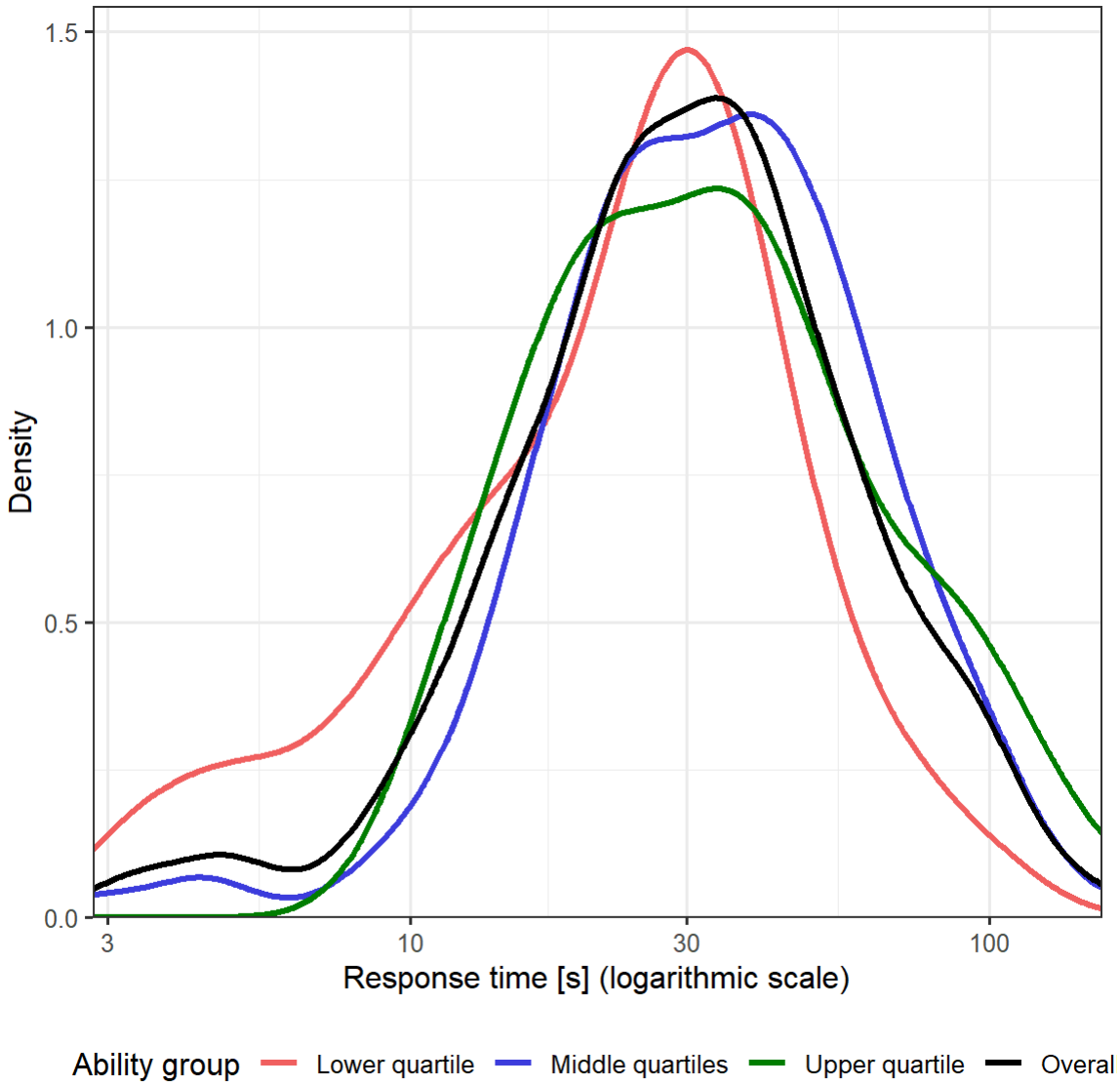
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Siatka A	22	14	11	47	37.9%	20.3%	33.3%	29.4%
Siatka B	23	30	10	63	39.7%	43.5%	30.3%	39.4%
Siatka D	13	25	12	50	22.4%	36.2%	36.4%	31.2%
Overall	58	69	33	160	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



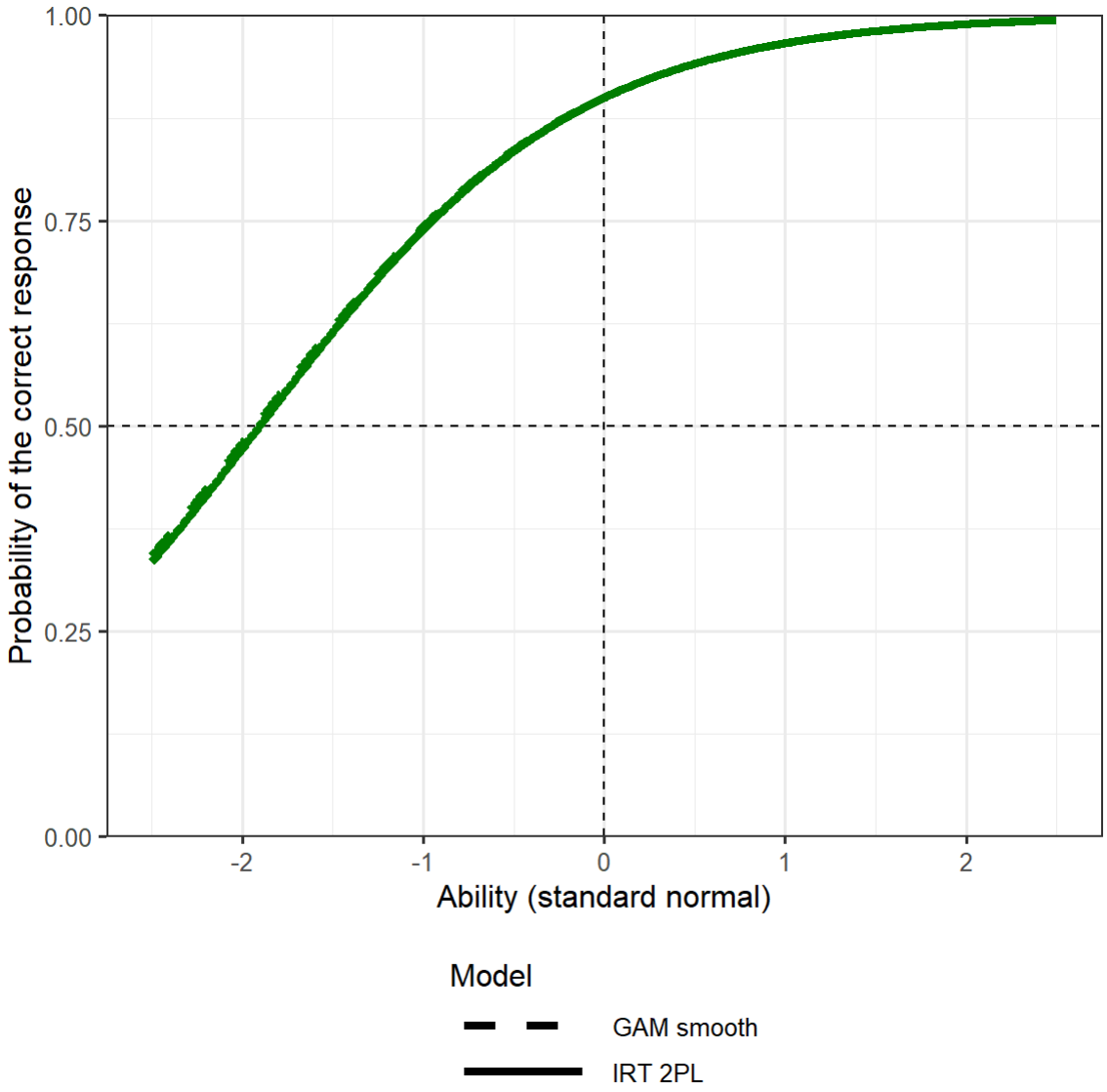
10.1.45.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.4	14.0	25.2	35.3	99.2	3.1	0.8
Middle quartiles	2.8	22.4	33.6	49.9	144.2	3.5	0.7
Upper quartile	10.1	20.2	33.6	51.4	156.3	3.5	0.7
Overall	2.8	20.2	30.5	46.0	156.3	3.4	0.7

10.1.46 M100004 (LLM)

10.1.46.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	85.23%	299	LLM simple setup	MC	0.36	0.21	1.15	−1.90	0.019

10.1.46.2 Gender DIF:

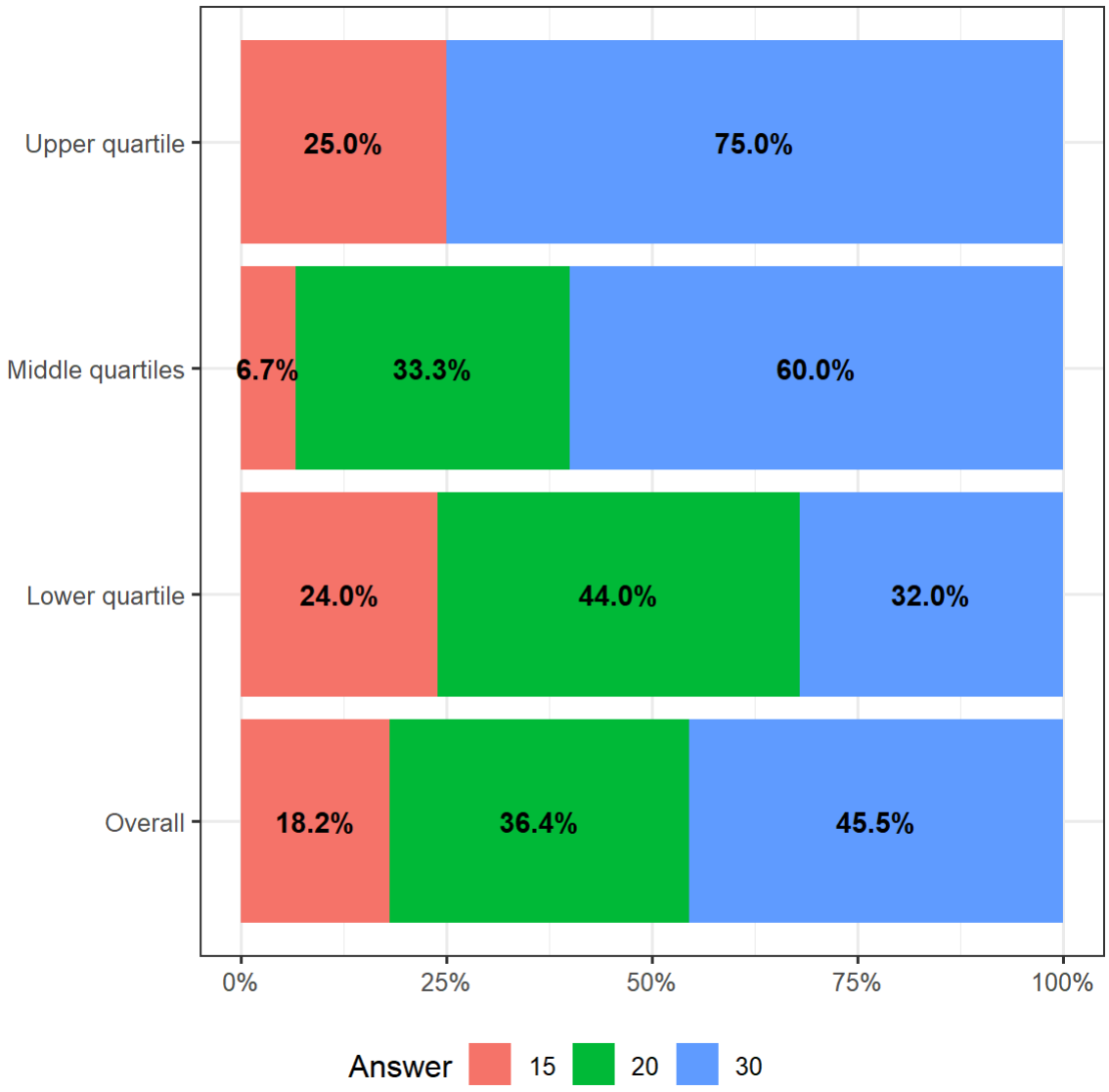
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.10	0.951	1.000	0.000	N/A

10.1.46.3 Frequency of answers:

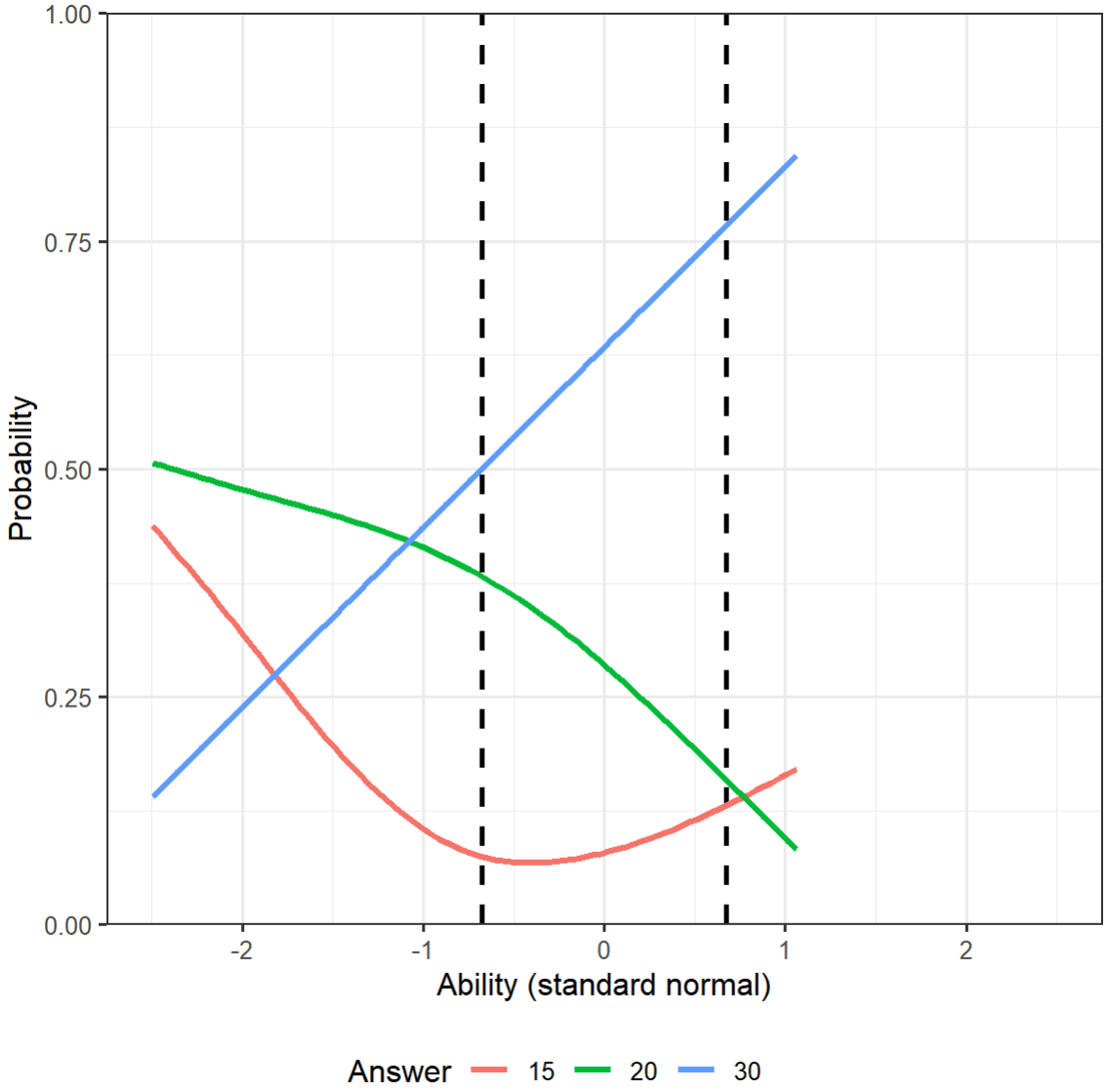
Answer	Frequency	Percent
15	8	2.7%
20	16	5.4%
25 (correct)	254	85.2%
30	20	6.7%
Overall	298	100.0%

10.1.46.4 Distractors functioning:

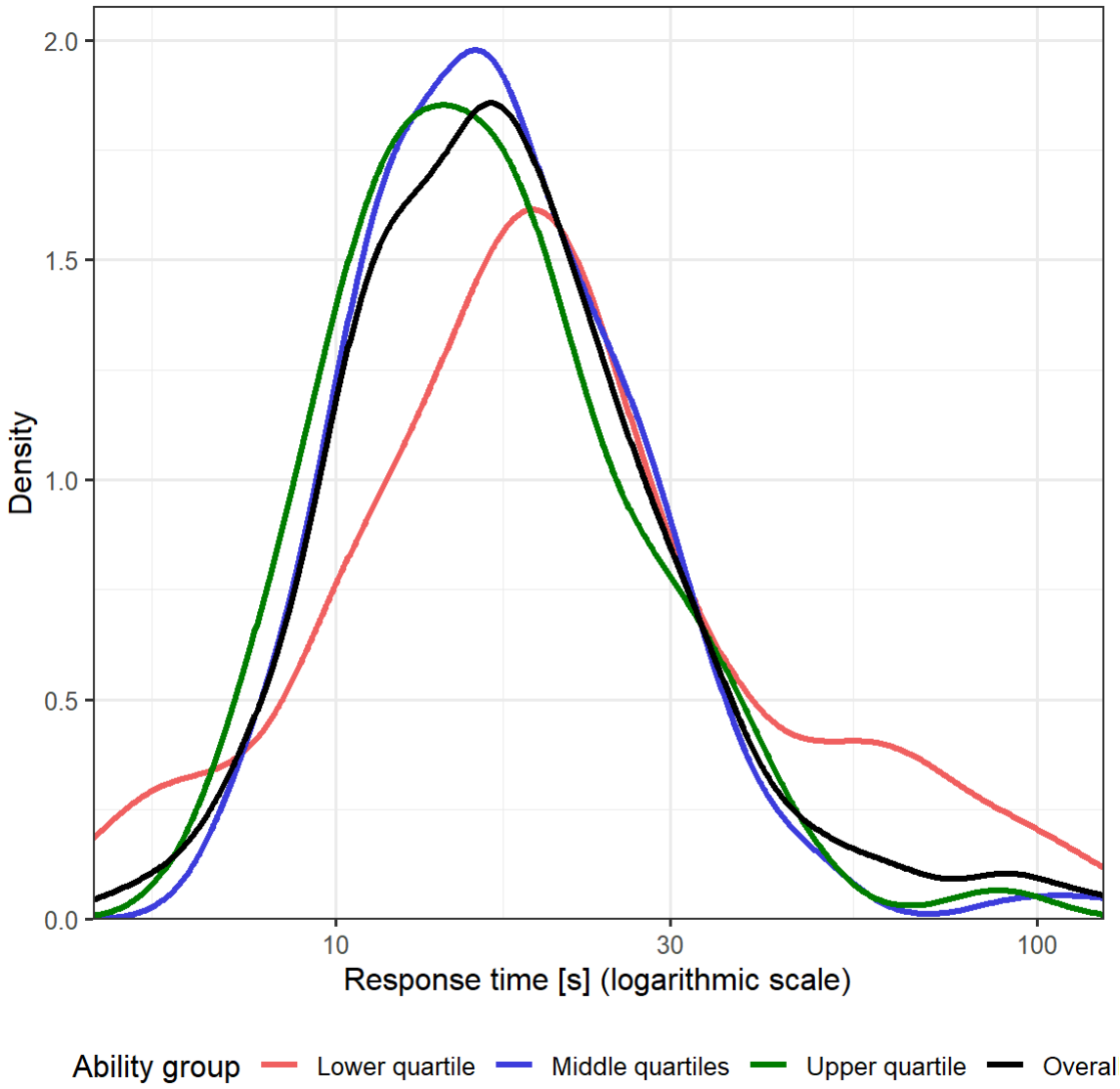
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
15	6	1	1	8	24.0%	6.7%	25.0%	18.2%
20	11	5	0	16	44.0%	33.3%	0.0%	36.4%
30	8	9	3	20	32.0%	60.0%	75.0%	45.5%
Overall	25	15	4	44	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



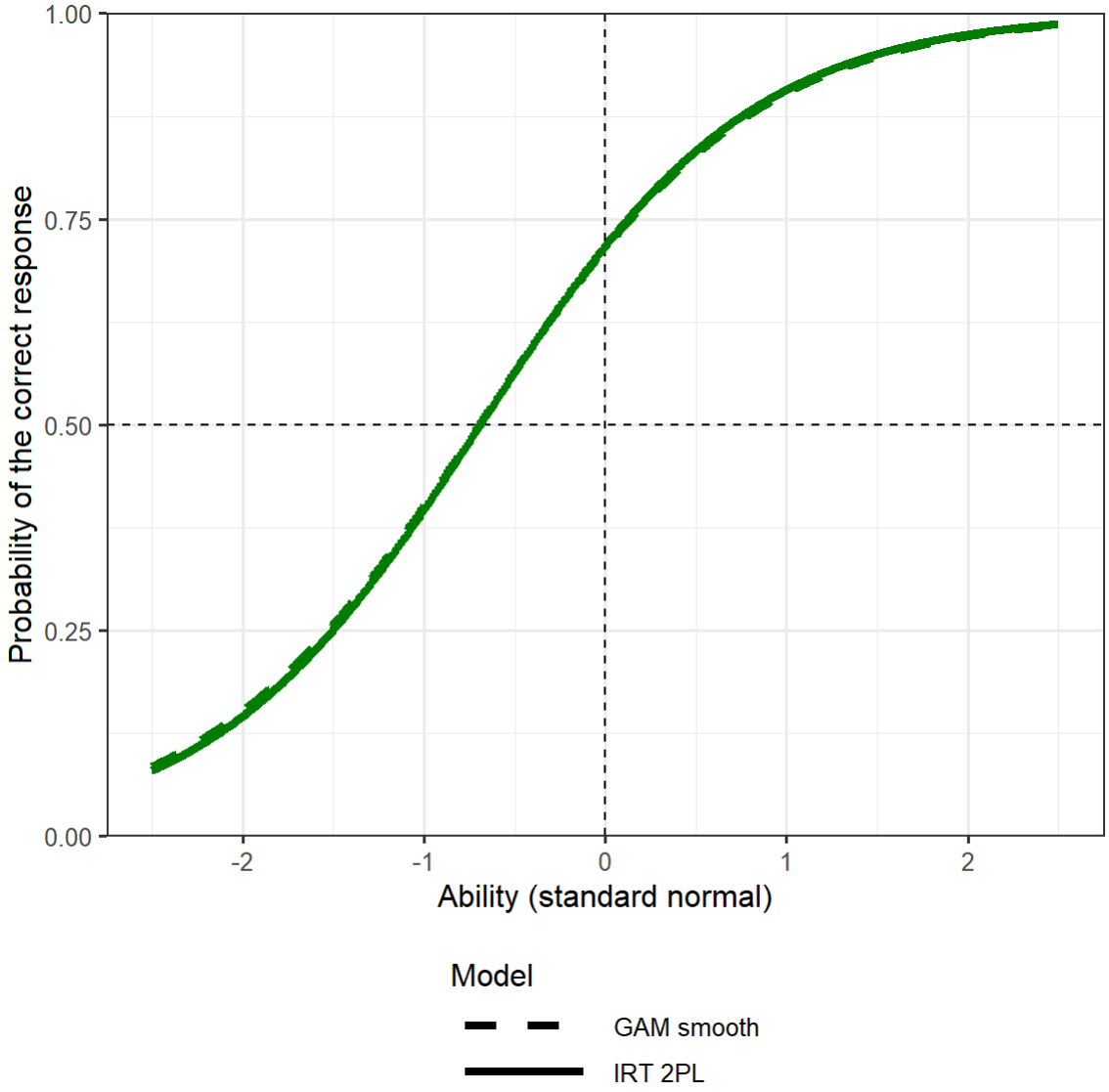
10.1.46.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.5	12.9	19.7	27.4	122.5	3.0	0.7
Middle quartiles	7.1	12.1	16.5	23.1	124.4	2.8	0.5
Upper quartile	6.8	11.5	15.3	21.8	88.4	2.8	0.5
Overall	4.5	12.1	16.8	23.5	124.4	2.9	0.5

10.1.47 M061185 (TIMSS)

10.1.47.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	65.57%	306	TIMSS	MC	0.49	0.48	1.35	−0.69	0.000

10.1.47.2 Gender DIF:

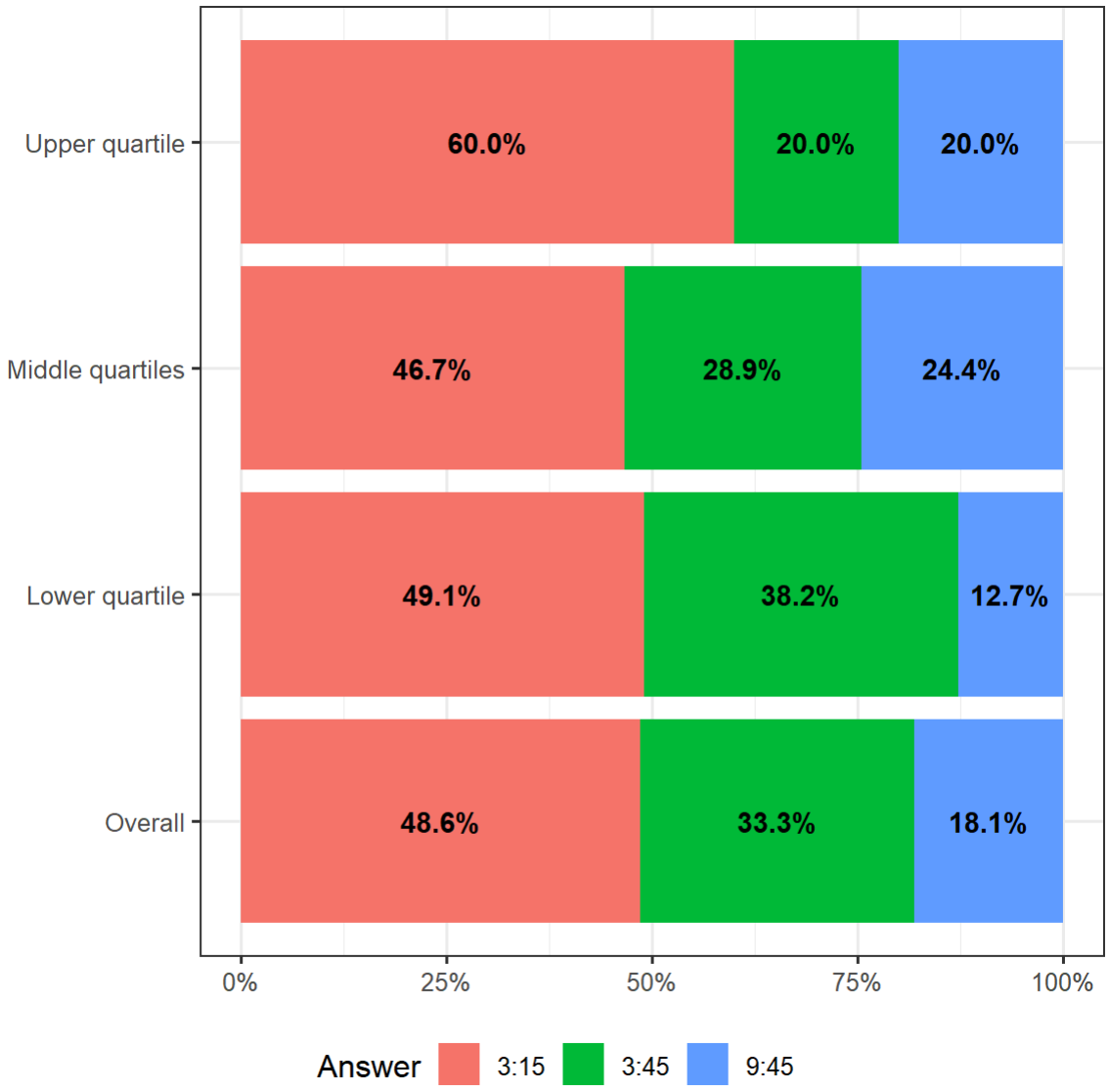
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.36	0.835	1.000	0.001	N/A

10.1.47.3 Frequency of answers:

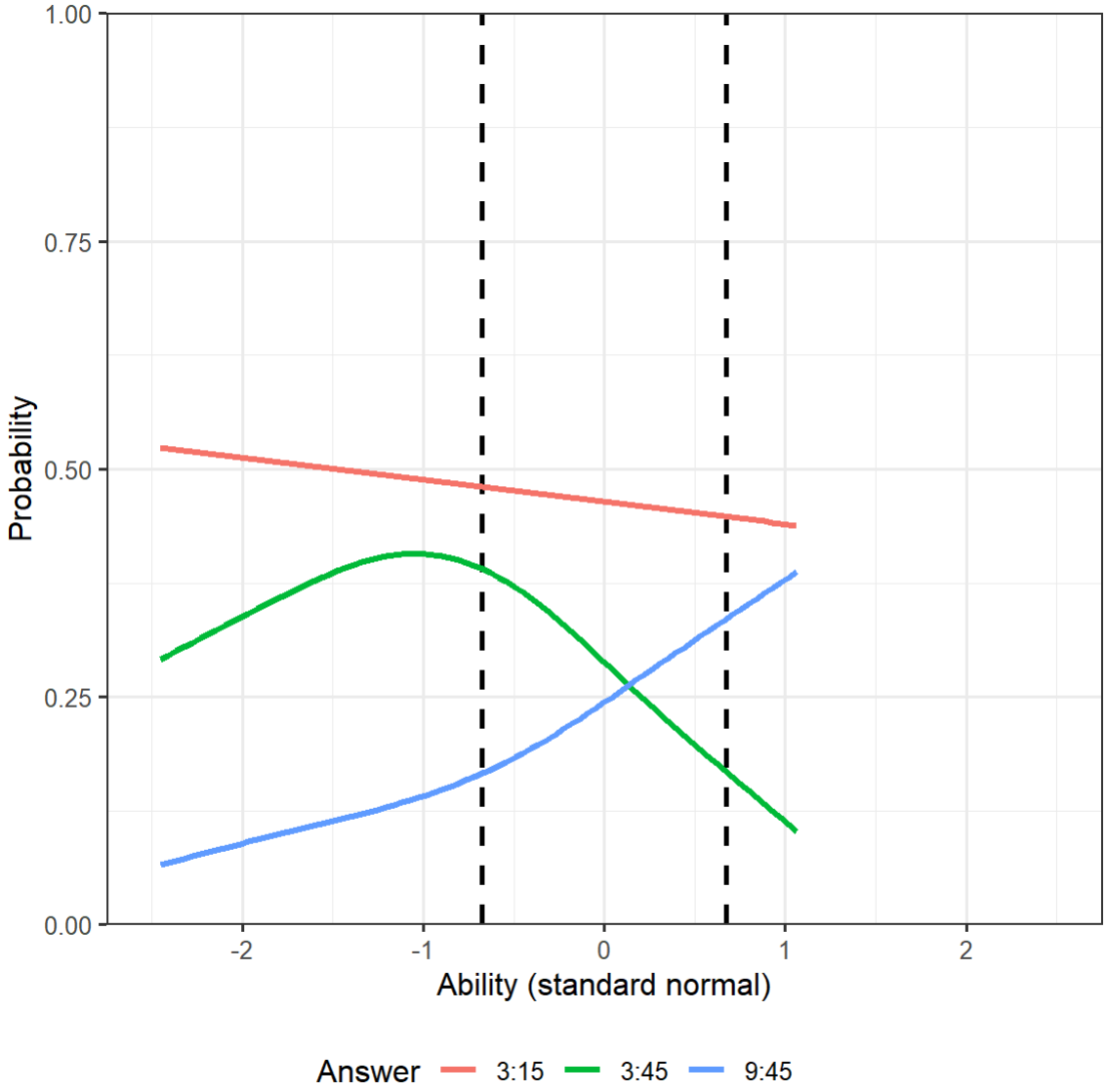
Answer	Frequency	Percent
3:15	51	16.7%
3:45	35	11.5%
9:00 (correct)	200	65.6%
9:45	19	6.2%
Overall	305	100.0%

10.1.47.4 Distractors functioning:

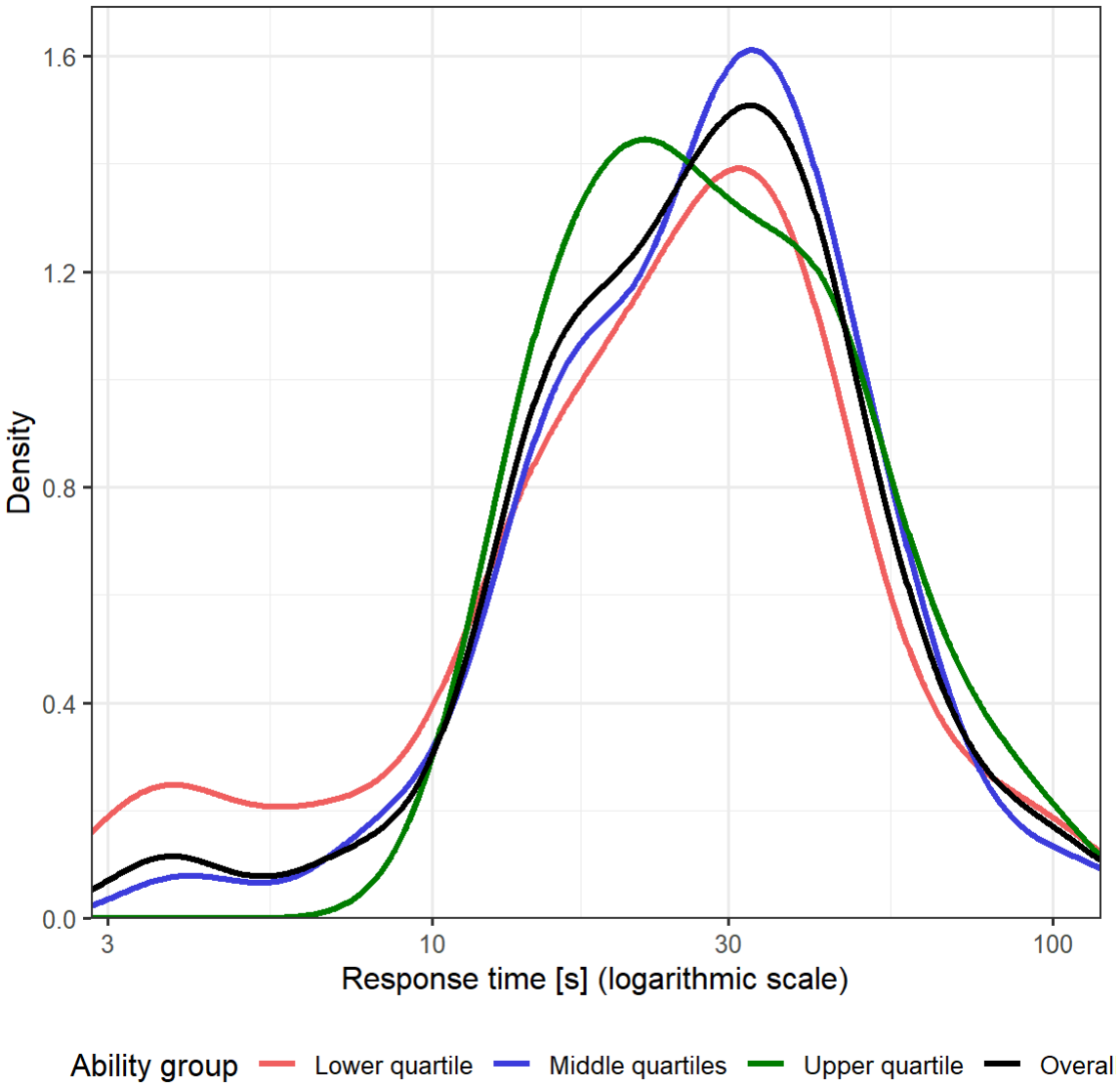
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
3:15	27	21	3	51	49.1%	46.7%	60.0%	48.6%
3:45	21	13	1	35	38.2%	28.9%	20.0%	33.3%
9:45	7	11	1	19	12.7%	24.4%	20.0%	18.1%
Overall	55	45	5	105	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



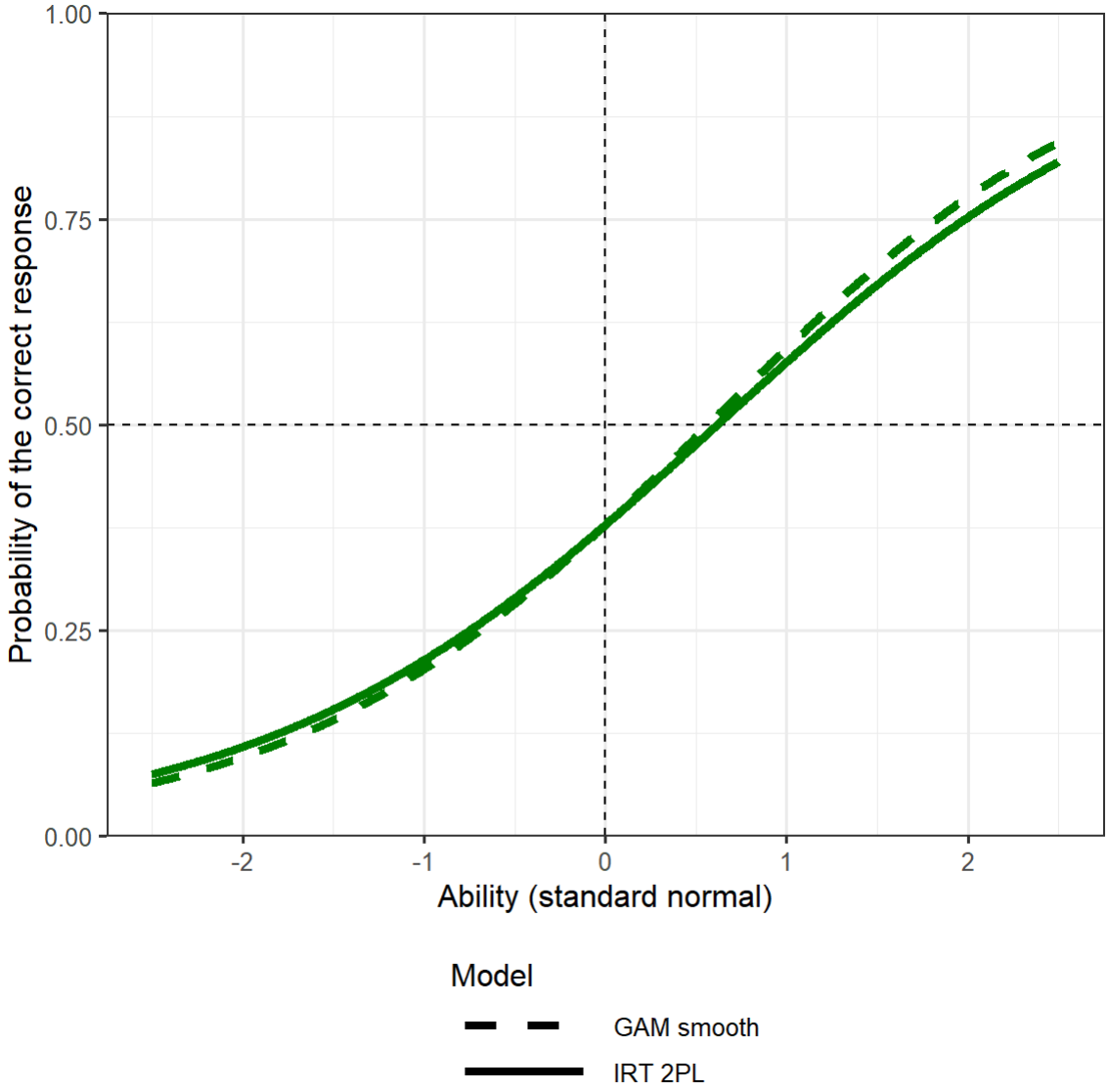
10.1.47.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.8	15.3	25.3	36.9	119.1	3.1	0.8
Middle quartiles	3.6	18.0	28.9	40.2	119.1	3.3	0.6
Upper quartile	10.2	18.6	27.7	42.7	112.5	3.3	0.5
Overall	2.8	17.3	28.1	40.0	119.1	3.3	0.7

10.1.48 M051070 (TIMSS)

10.1.48.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	40.14%	294	TIMSS	MC	0.37	0.33	0.80	0.62	0.000

10.1.48.2 Gender DIF:

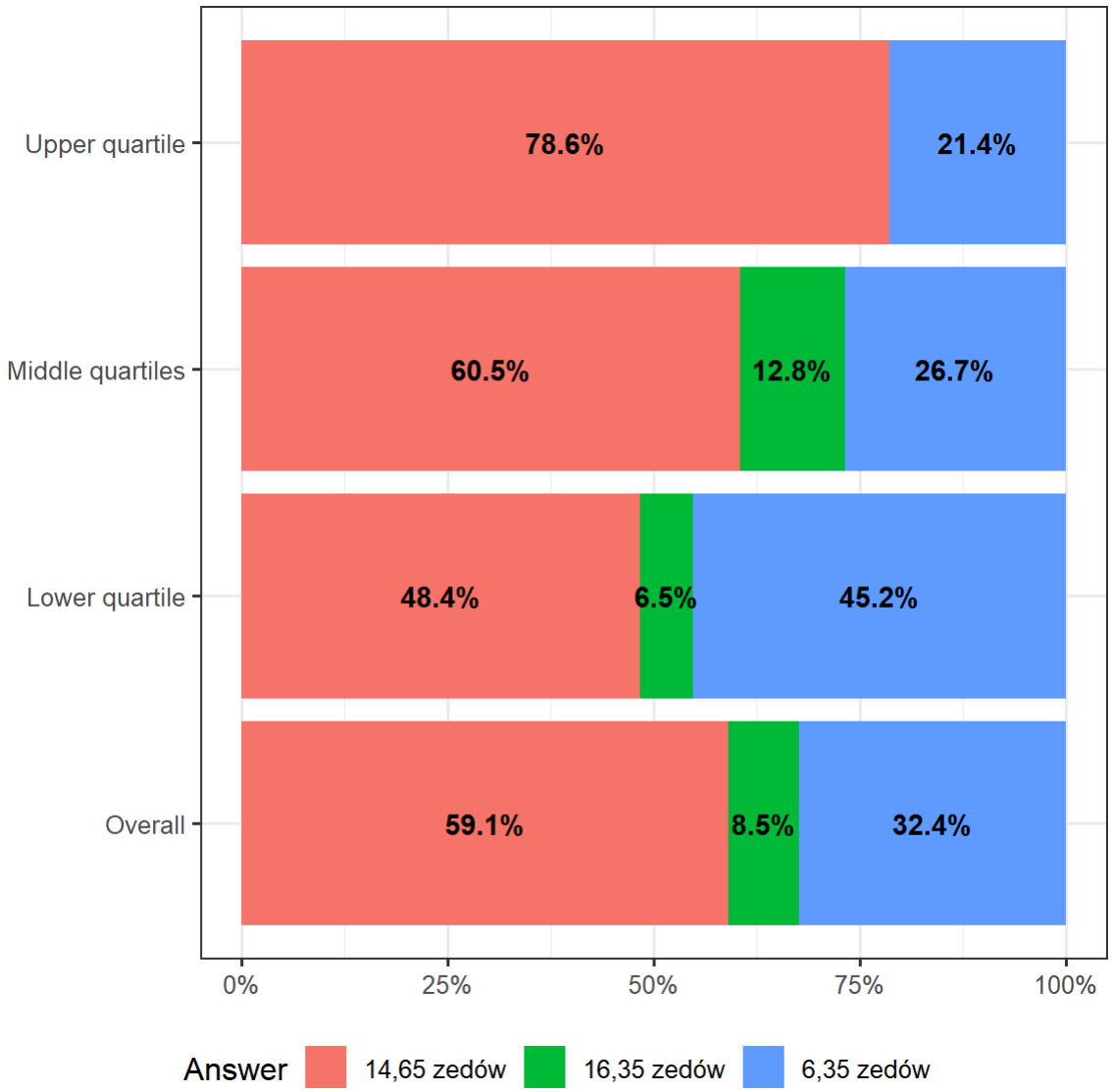
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.71	0.258	1.000	0.008	N/A

10.1.48.3 Frequency of answers:

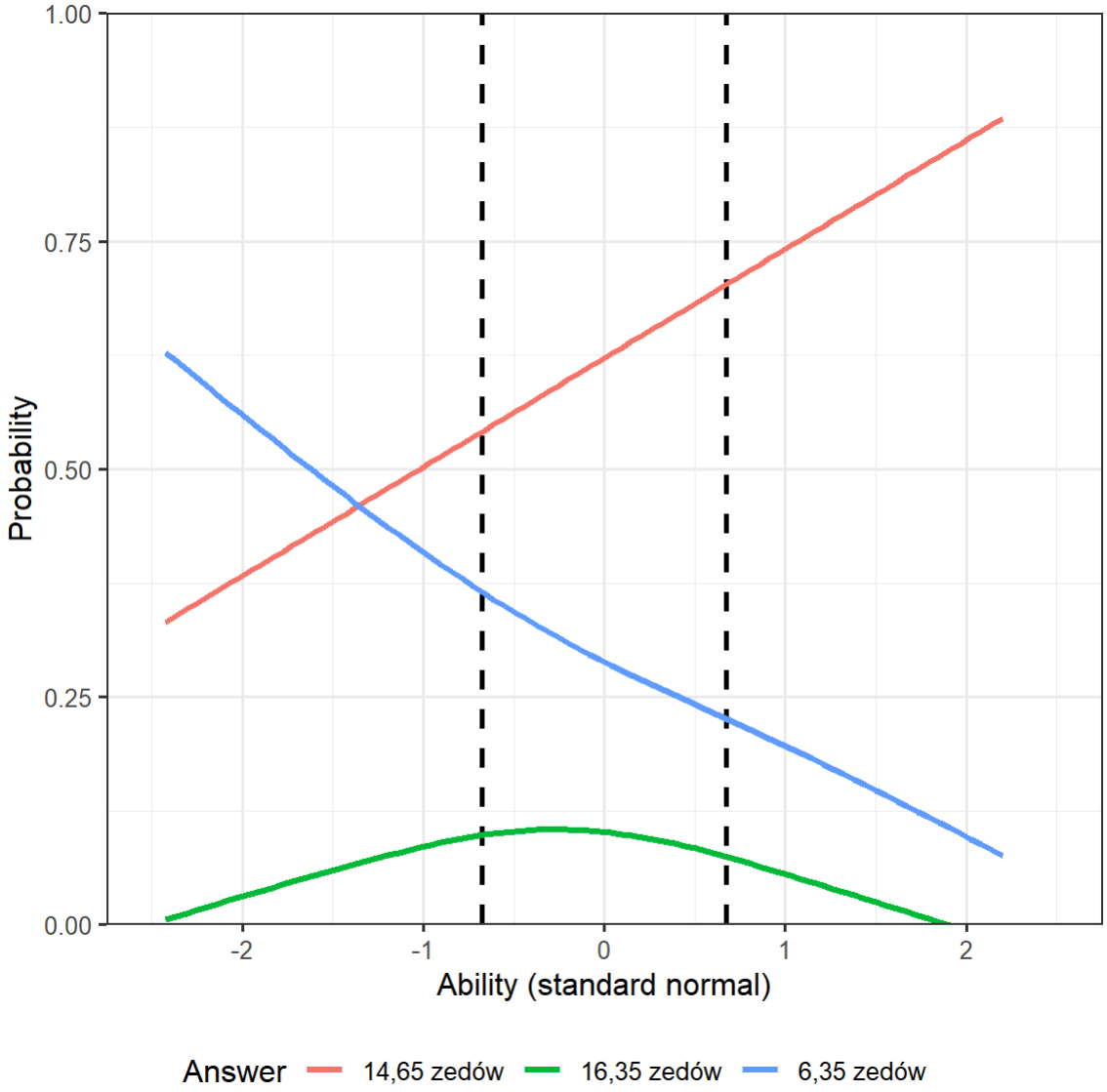
Answer	Frequency	Percent
13,65 zedów (correct)	118	40.1%
14,65 zedów	104	35.4%
16,35 zedów	15	5.1%
6,35 zedów	57	19.4%
Overall	294	100.0%

10.1.48.4 Distractors functioning:

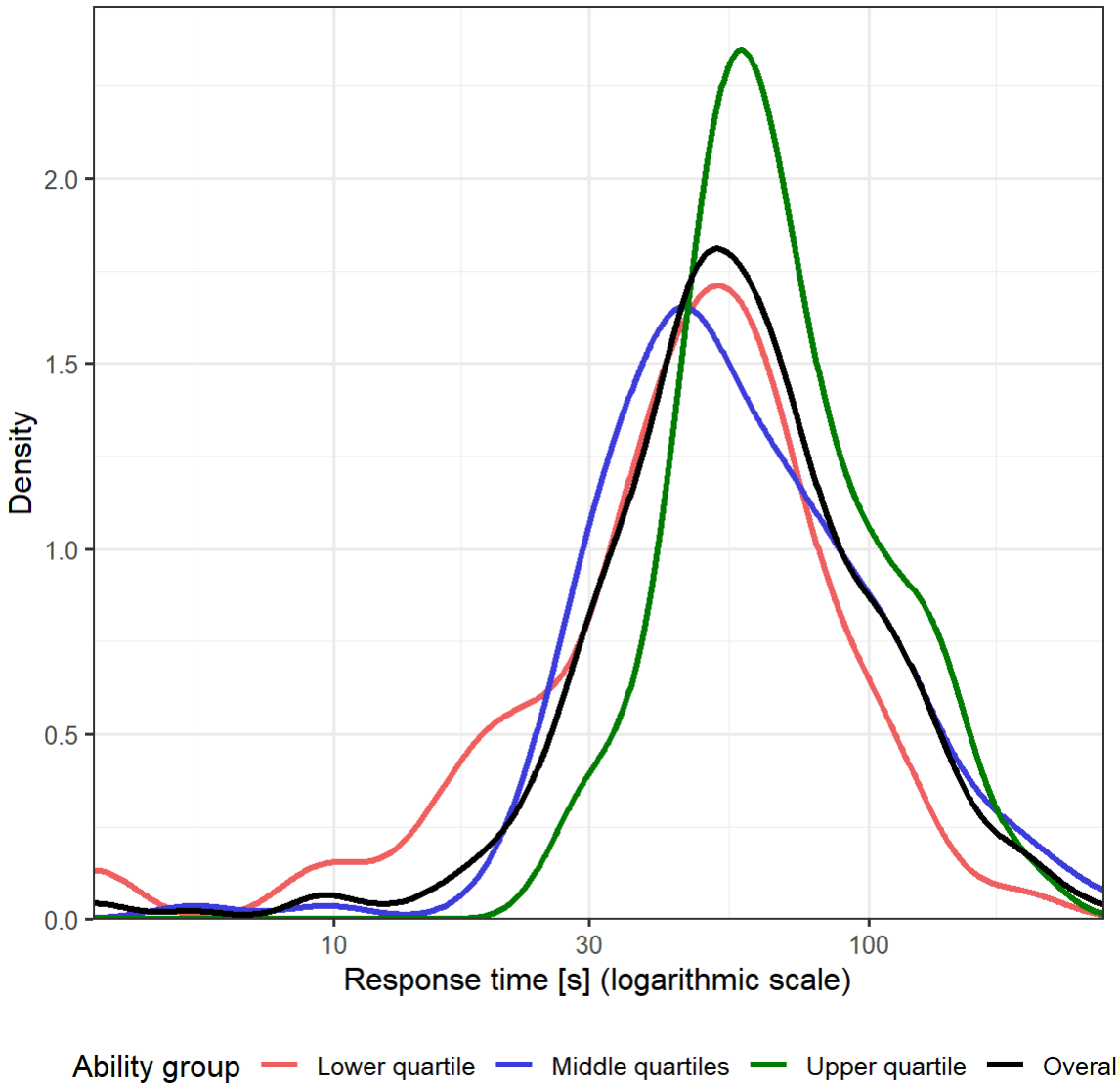
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
14,65 zedów	30	52	22	104	48.4%	60.5%	78.6%	59.1%
16,35 zedów	4	11	0	15	6.5%	12.8%	0.0%	8.5%
6,35 zedów	28	23	6	57	45.2%	26.7%	21.4%	32.4%
Overall	62	86	28	176	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



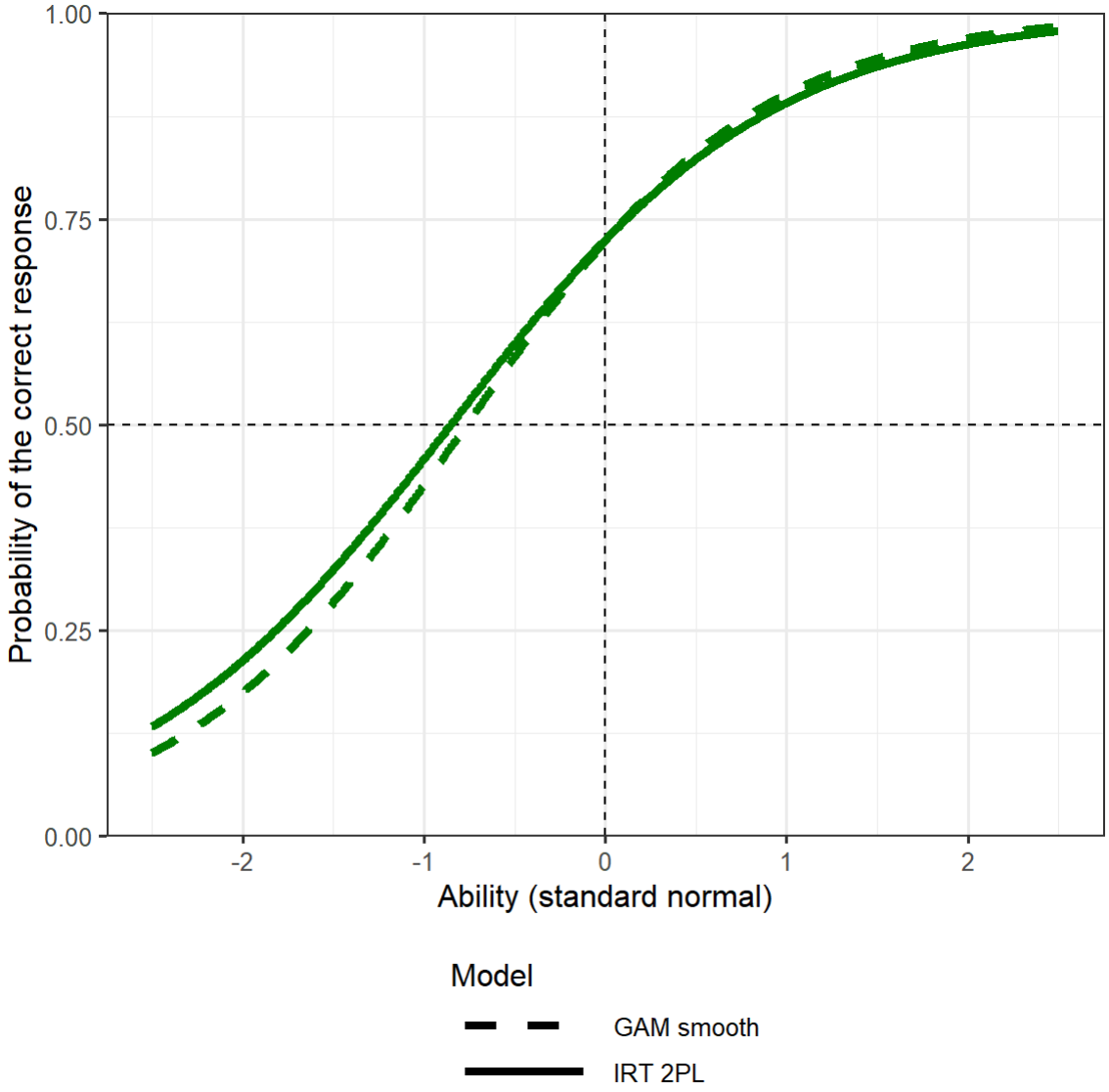
10.1.48.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	33.8	48.2	65.2	194.0	3.8	0.7
Middle quartiles	5.5	38.8	51.5	82.1	274.2	4.0	0.6
Upper quartile	28.2	51.6	62.8	91.0	204.8	4.2	0.4
Overall	3.5	40.0	55.6	79.7	274.2	4.0	0.6

10.1.49 M061050 (TIMSS)

10.1.49.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	67.76%	304	TIMSS	MC	0.46	0.41	1.14	−0.85	0.000

10.1.49.2 Gender DIF:

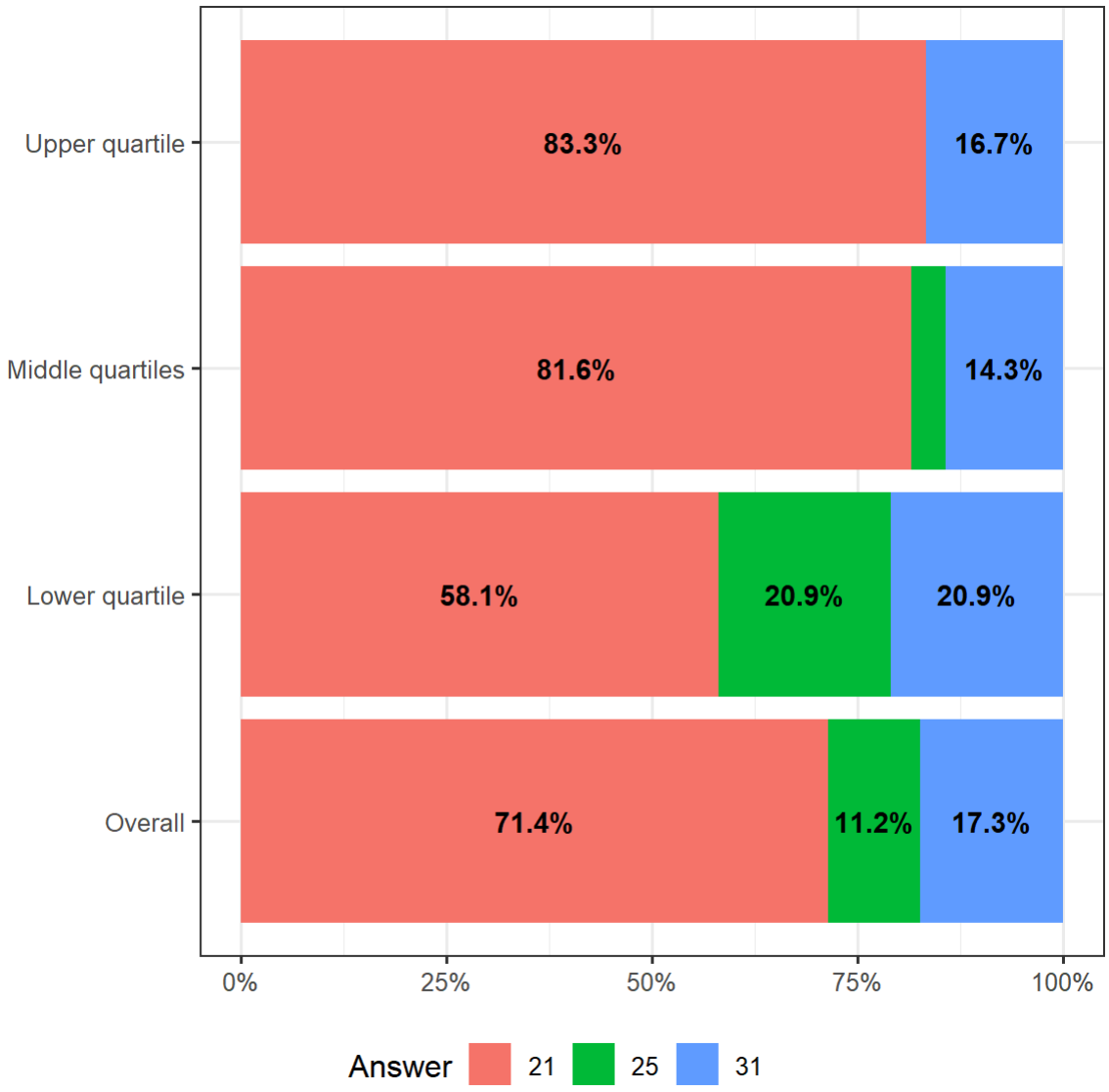
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.95	0.377	1.000	0.006	N/A

10.1.49.3 Frequency of answers:

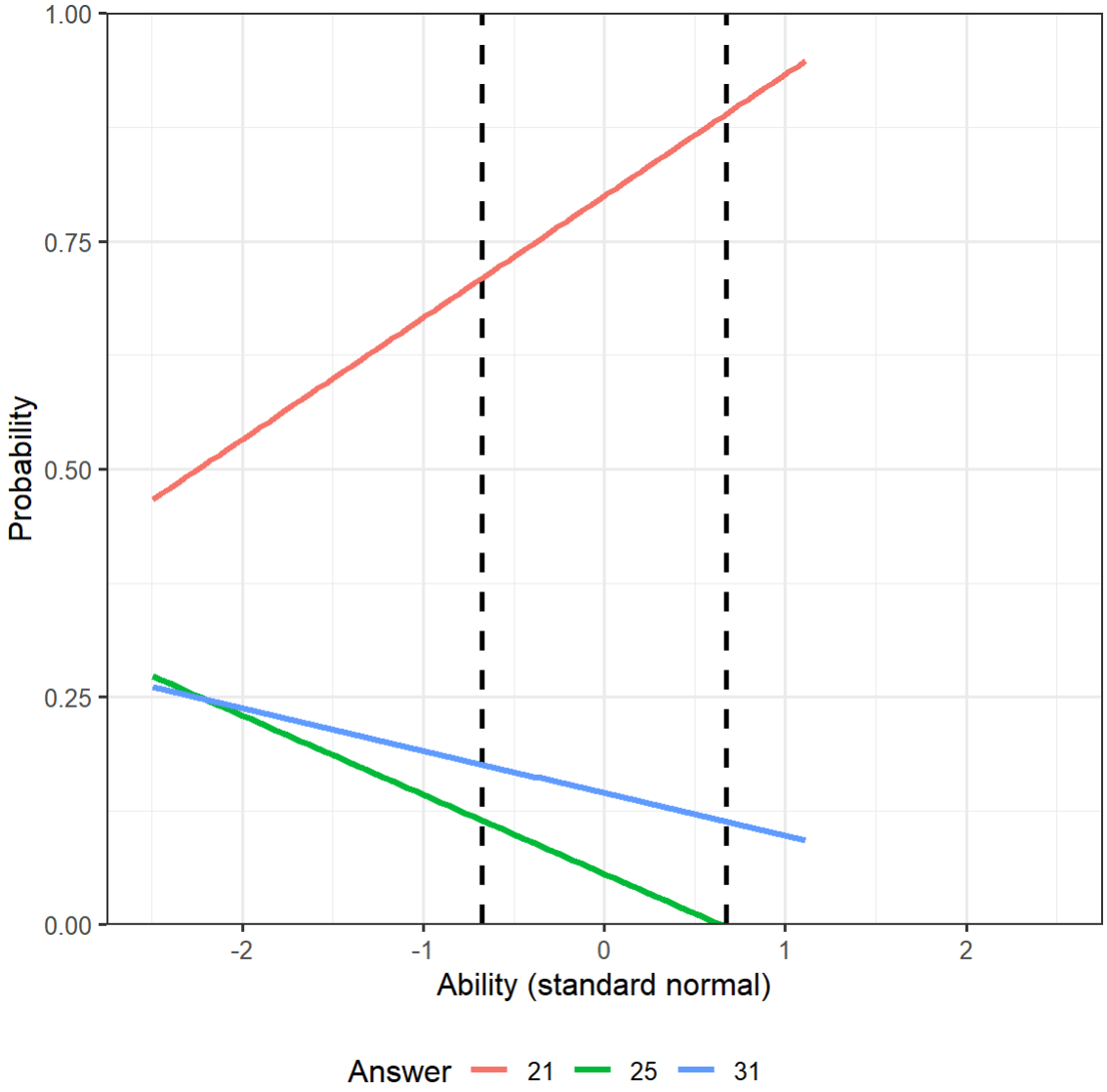
Answer	Frequency	Percent
11 (correct)	206	67.8%
21	70	23.0%
25	11	3.6%
31	17	5.6%
Overall	304	100.0%

10.1.49.4 Distractors functioning:

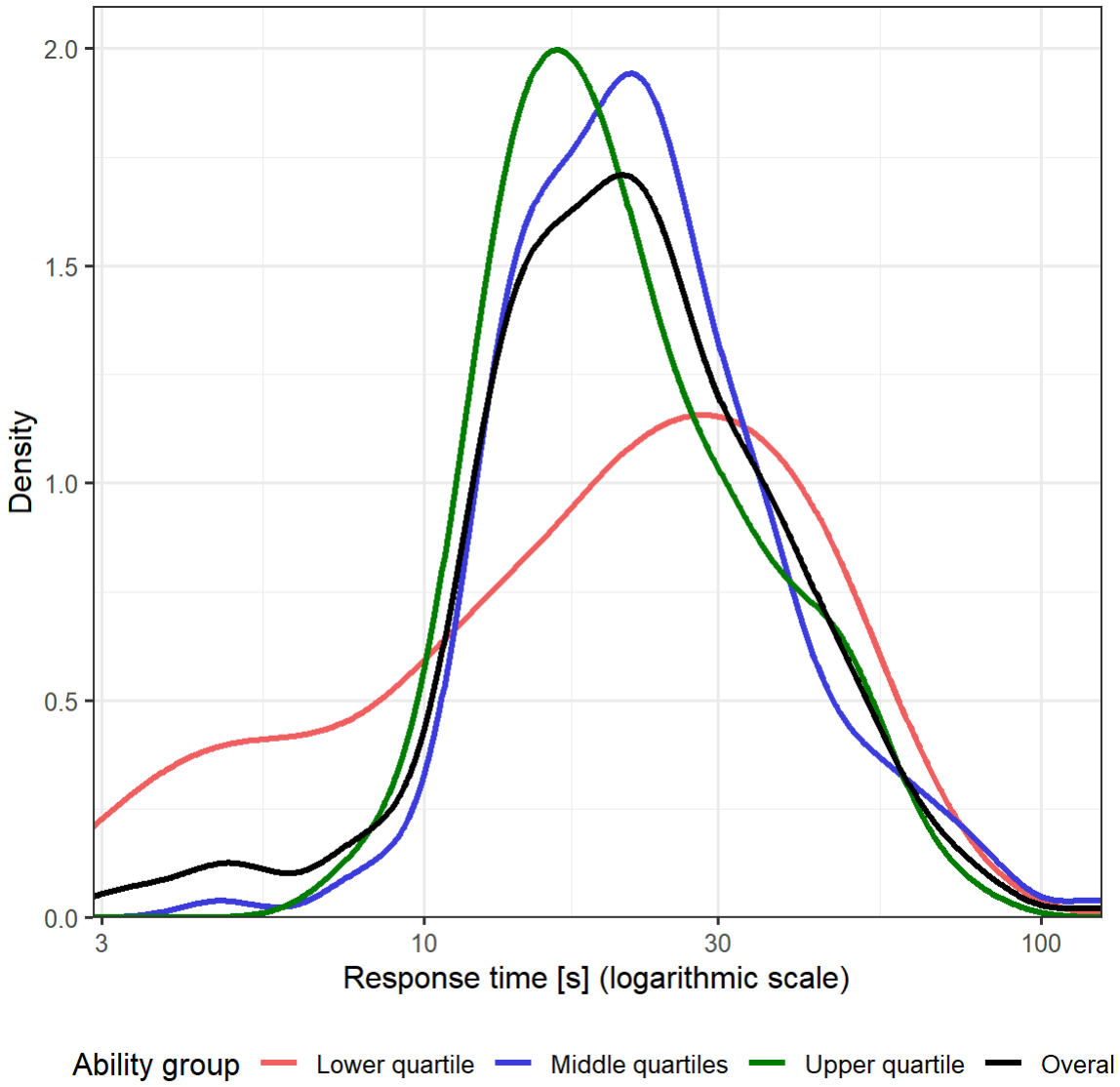
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
21	25	40	5	70	58.1%	81.6%	83.3%	71.4%
25	9	2	0	11	20.9%	4.1%	0.0%	11.2%
31	9	7	1	17	20.9%	14.3%	16.7%	17.3%
Overall	43	49	6	98	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



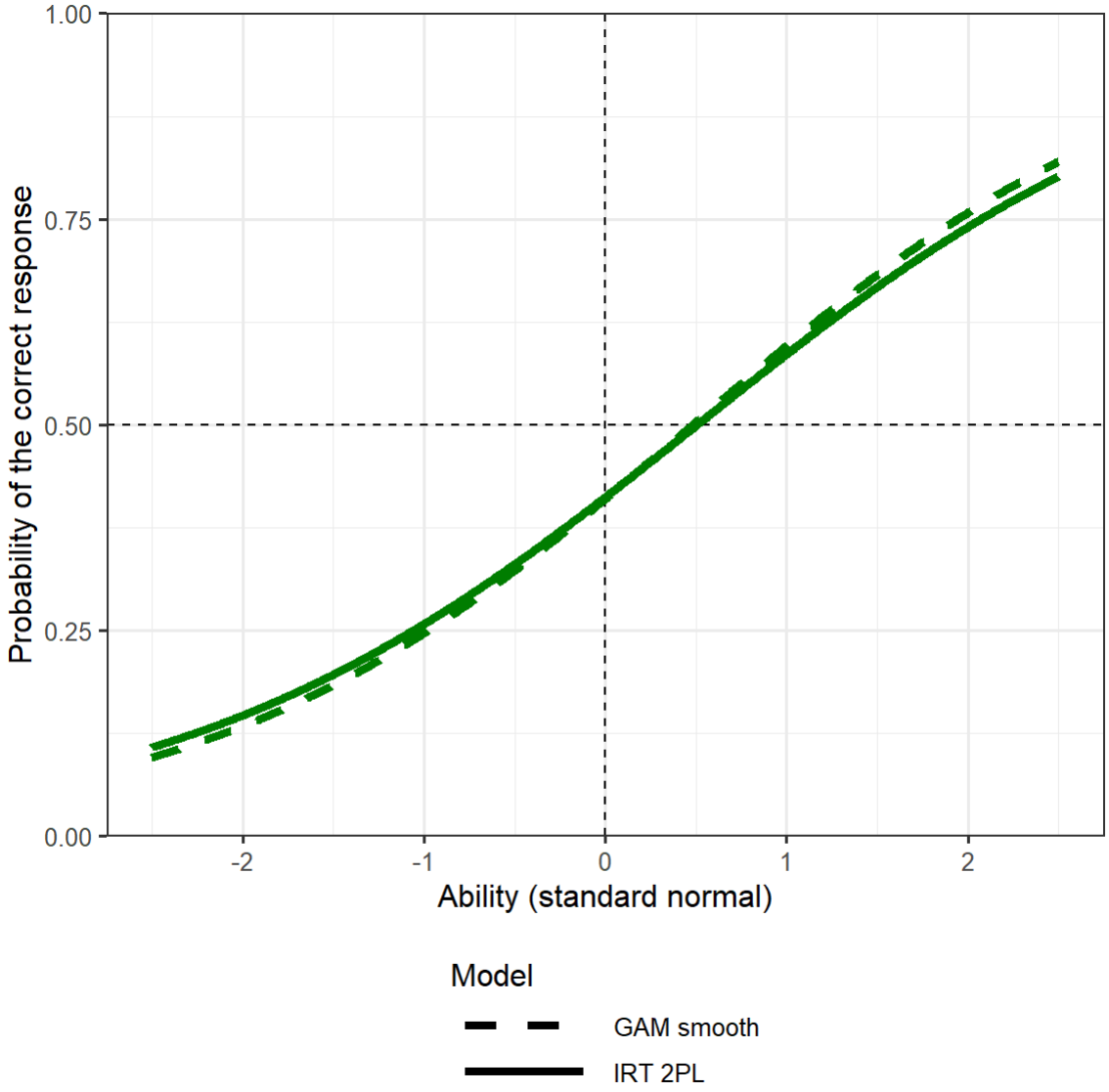
10.1.49.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.9	11.2	22.6	34.9	62.2	2.9	0.8
Middle quartiles	4.7	15.8	22.2	29.3	125.4	3.1	0.5
Upper quartile	7.8	14.5	18.9	28.5	71.7	3.0	0.5
Overall	2.9	14.9	21.2	30.6	125.4	3.0	0.6

10.1.50 M041096 (TIMSS)

10.1.50.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	41.31%	306	TIMSS	MC	0.33	0.26	0.70	0.51	0.000

10.1.50.2 Gender DIF:

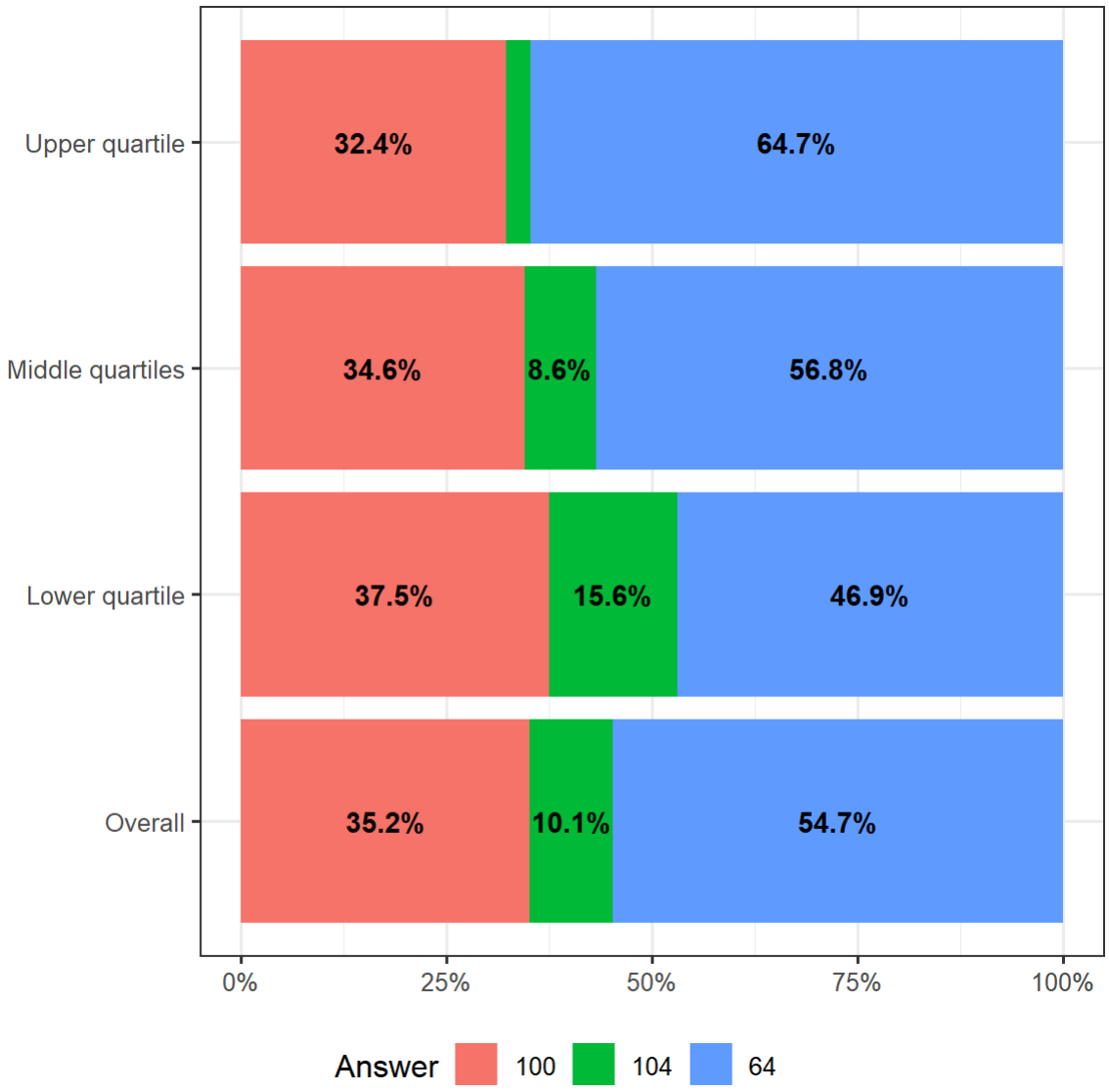
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.30	0.860	1.000	0.001	N/A

10.1.50.3 Frequency of answers:

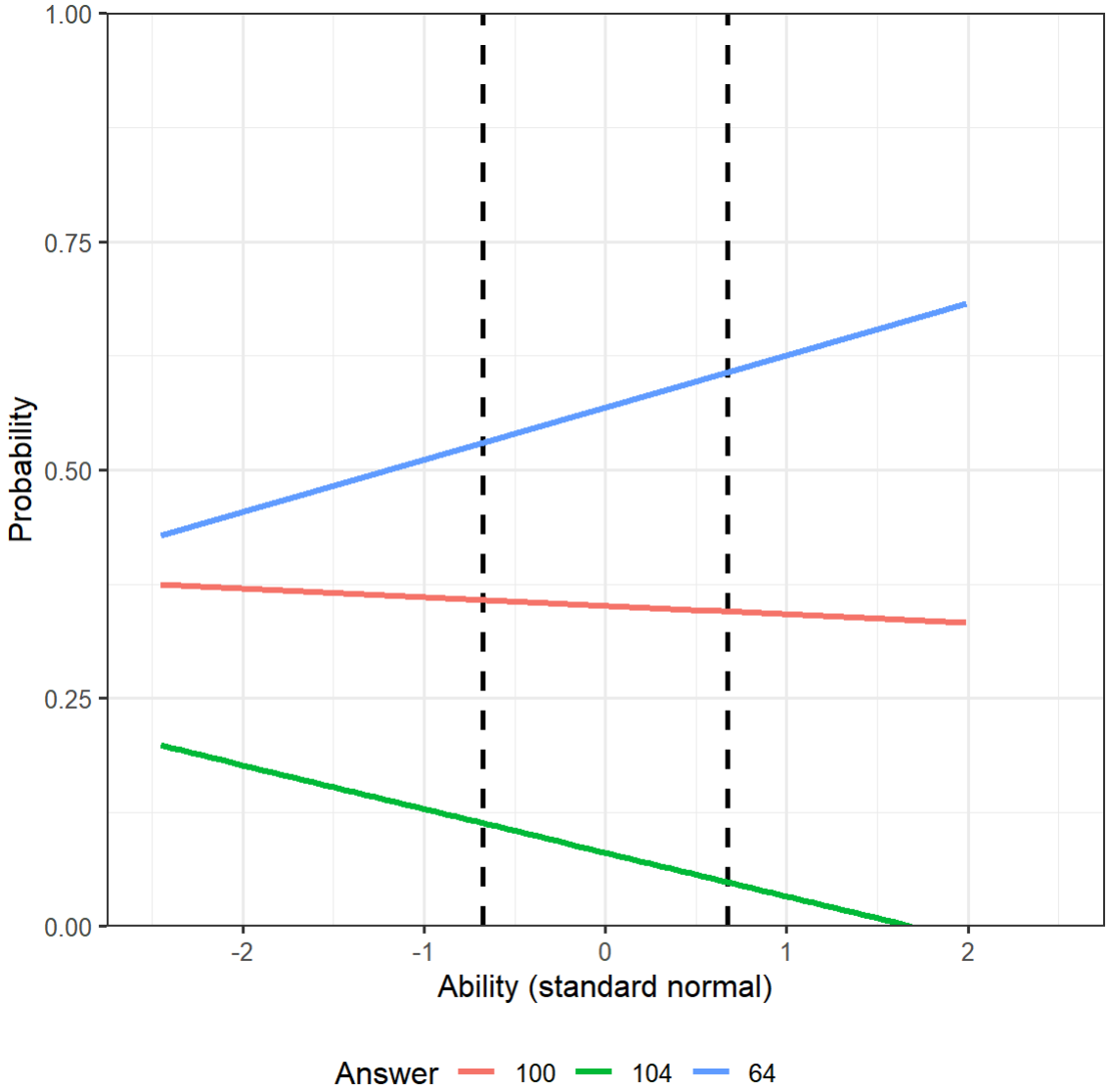
Answer	Frequency	Percent
64	98	32.1%
100	63	20.7%
104	18	5.9%
110 (correct)	126	41.3%
Overall	305	100.0%

10.1.50.4 Distractors functioning:

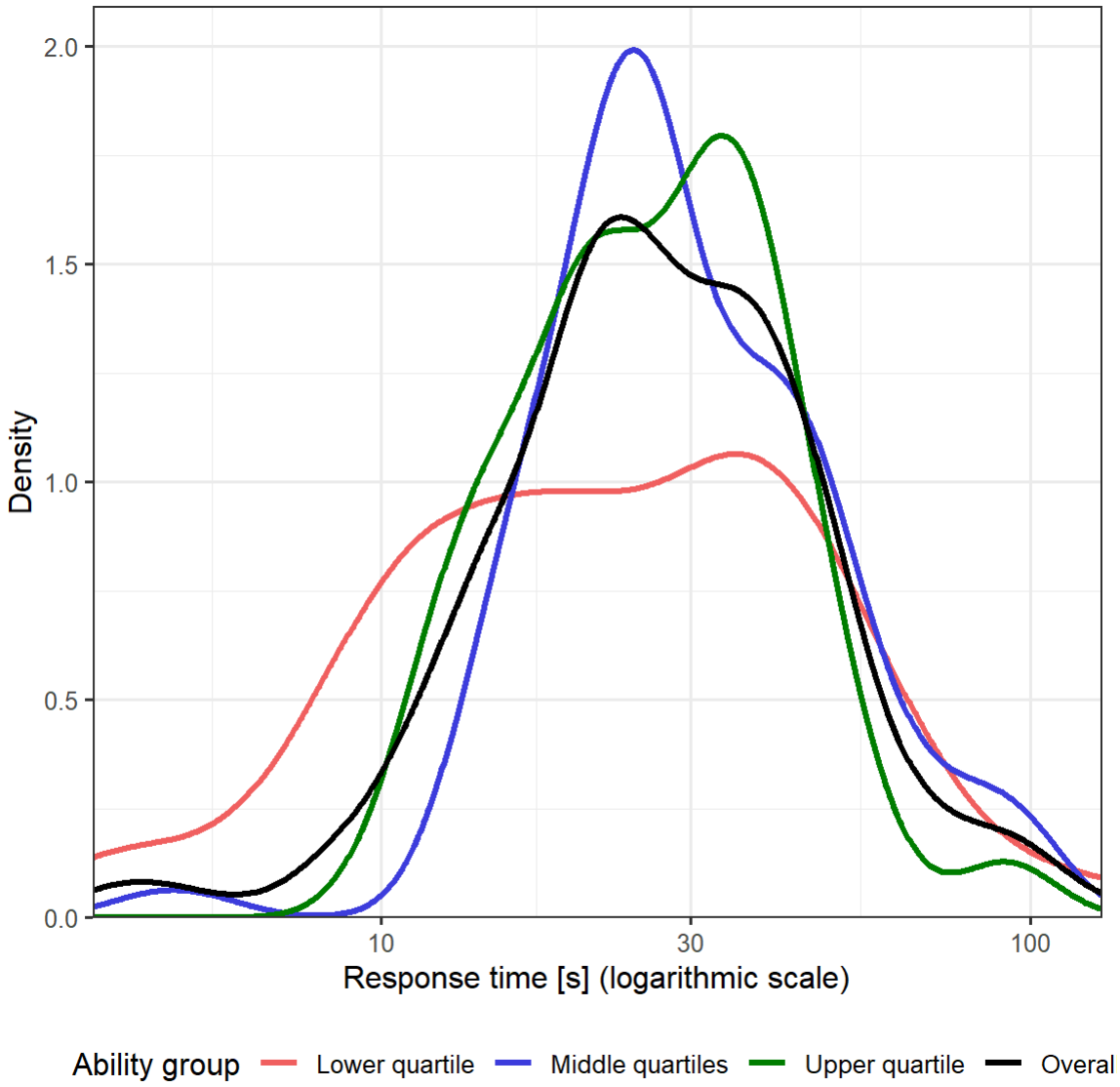
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
100	24	28	11	63	37.5%	34.6%	32.4%	35.2%
104	10	7	1	18	15.6%	8.6%	2.9%	10.1%
64	30	46	22	98	46.9%	56.8%	64.7%	54.7%
Overall	64	81	34	179	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



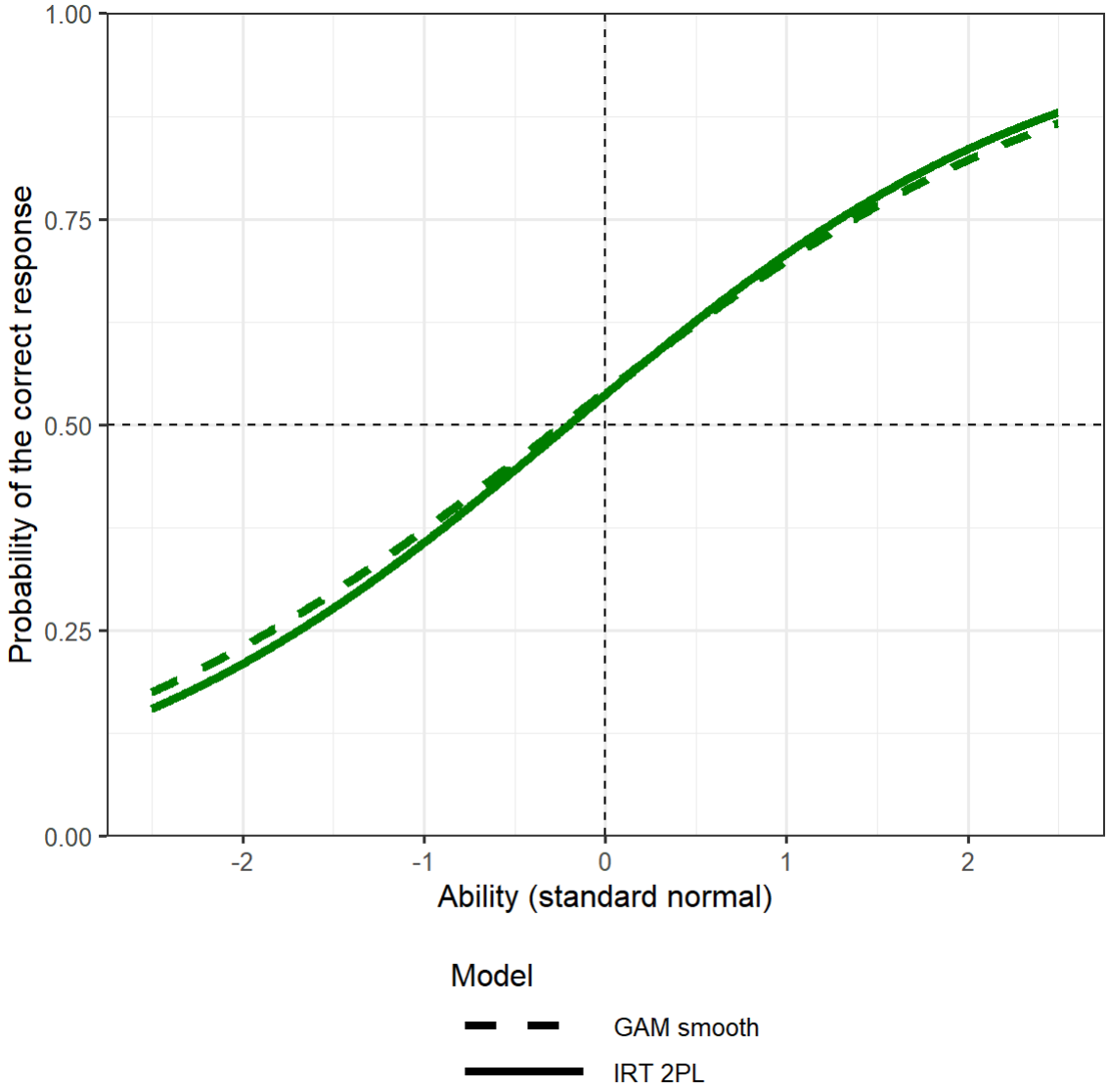
10.1.50.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.6	12.6	20.9	37.8	128.7	3.1	0.8
Middle quartiles	4.3	21.2	26.7	41.8	105.6	3.4	0.5
Upper quartile	10.3	19.3	27.6	36.2	94.7	3.3	0.5
Overall	3.6	19.1	26.0	38.5	128.7	3.3	0.6

10.1.51 M051093 (TIMSS)

10.1.51.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	54.08%	294	TIMSS	MC	0.32	0.28	0.74	−0.20	0.012

10.1.51.2 Gender DIF:

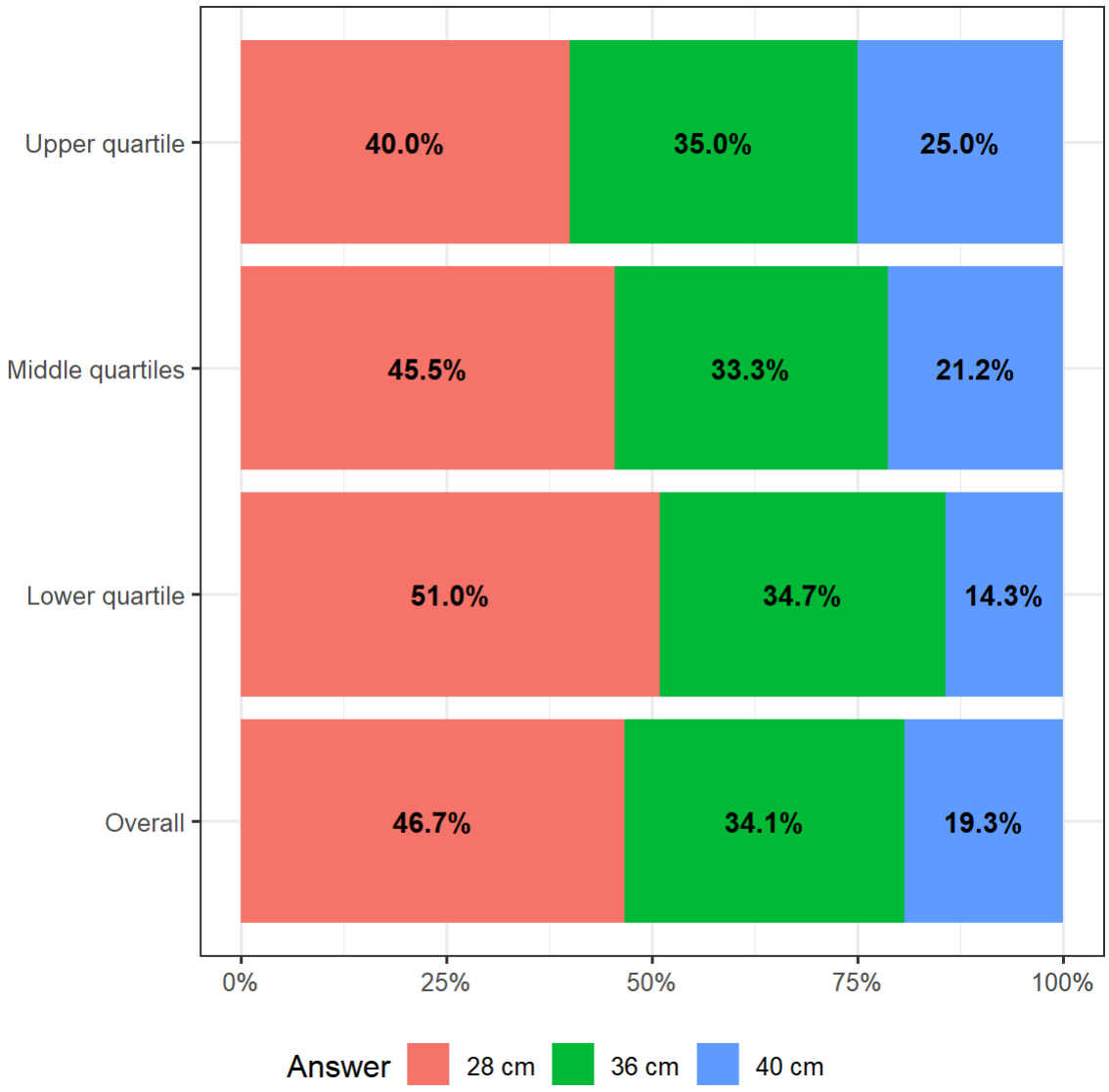
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.08	0.961	1.000	0.000	N/A

10.1.51.3 Frequency of answers:

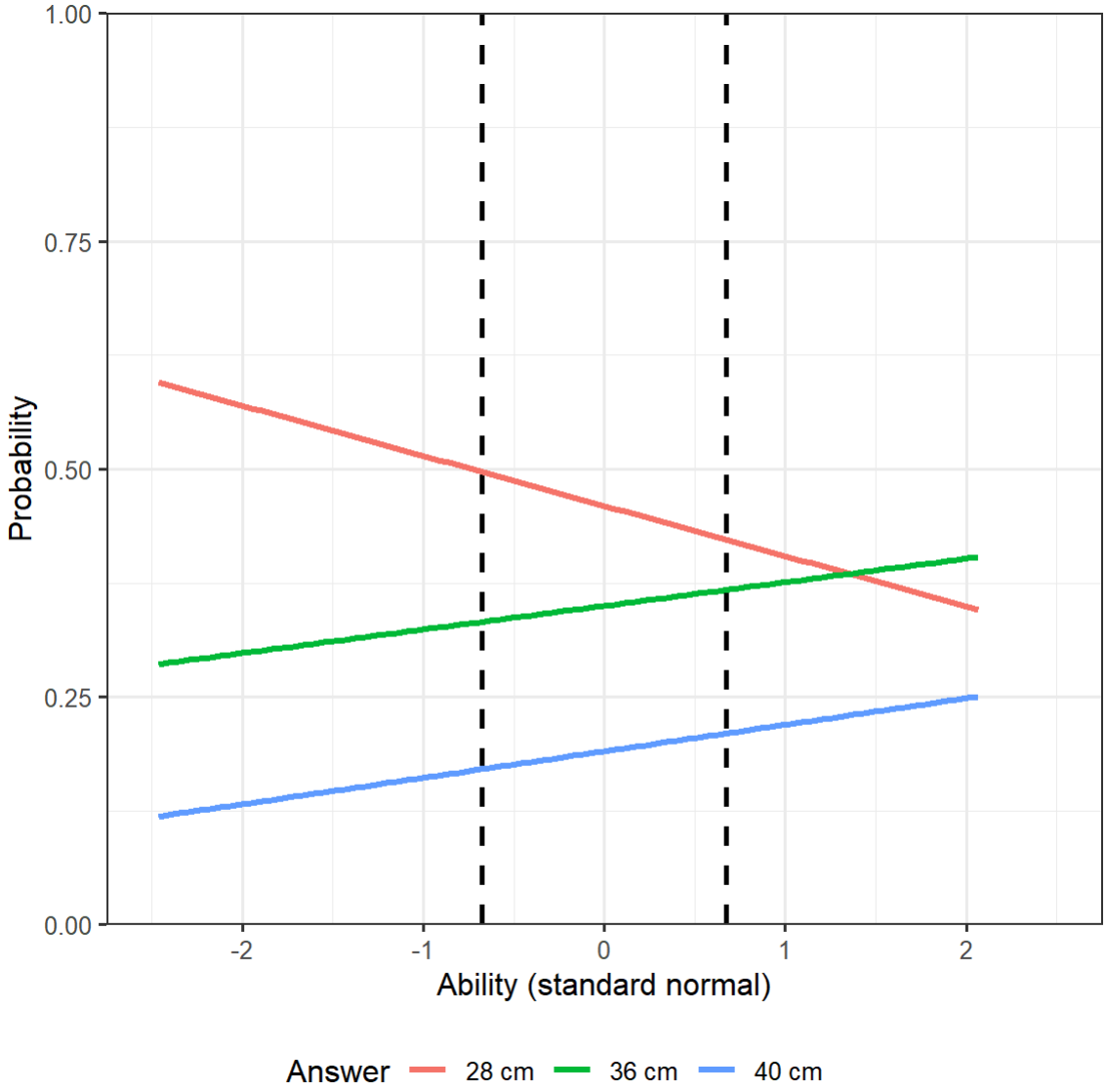
Answer	Frequency	Percent
28 cm	63	21.4%
32 cm (correct)	159	54.1%
36 cm	46	15.6%
40 cm	26	8.8%
Overall	294	100.0%

10.1.51.4 Distractors functioning:

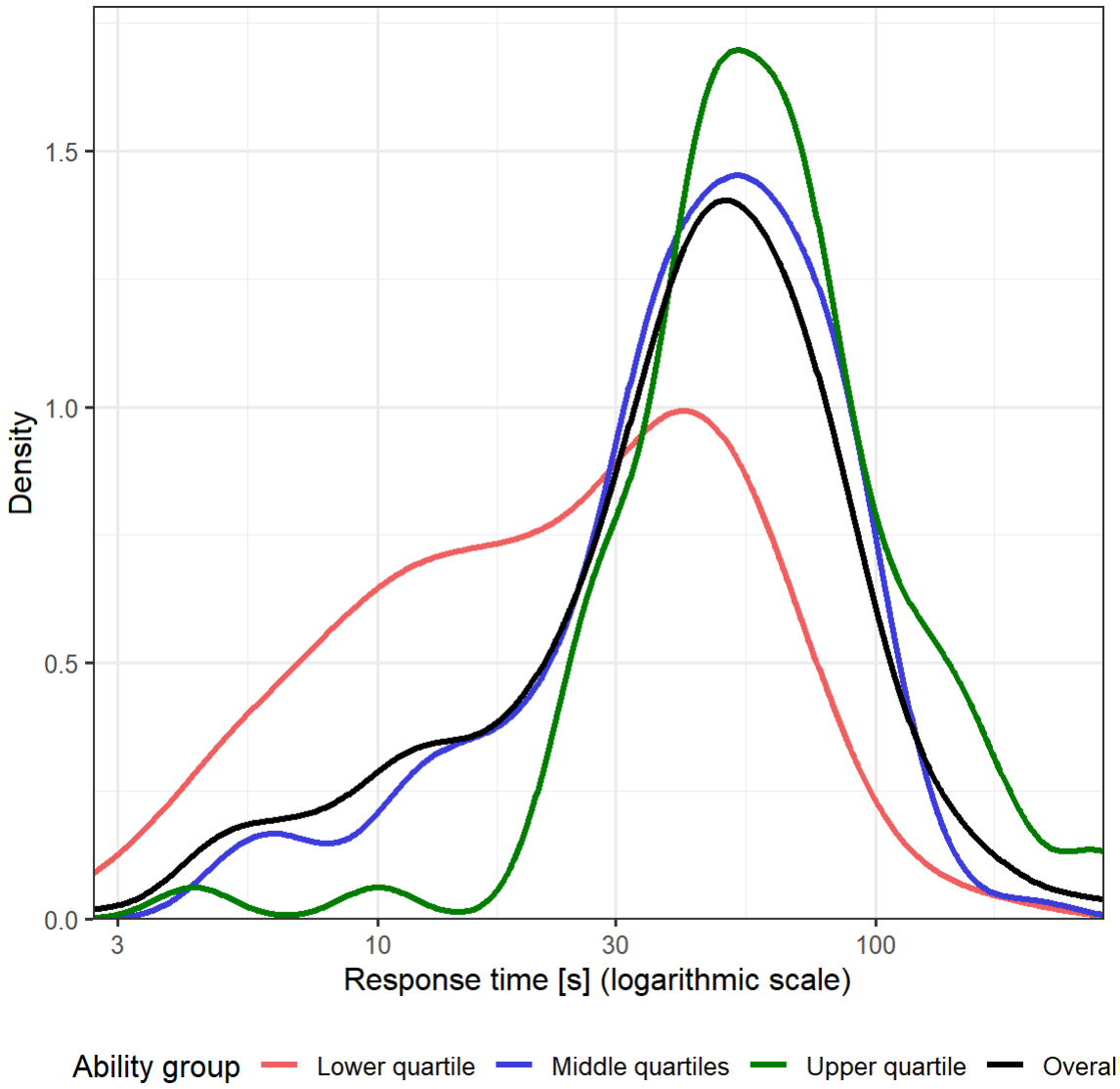
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
28 cm	25	30	8	63	51.0%	45.5%	40.0%	46.7%
36 cm	17	22	7	46	34.7%	33.3%	35.0%	34.1%
40 cm	7	14	5	26	14.3%	21.2%	25.0%	19.3%
Overall	49	66	20	135	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 12 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



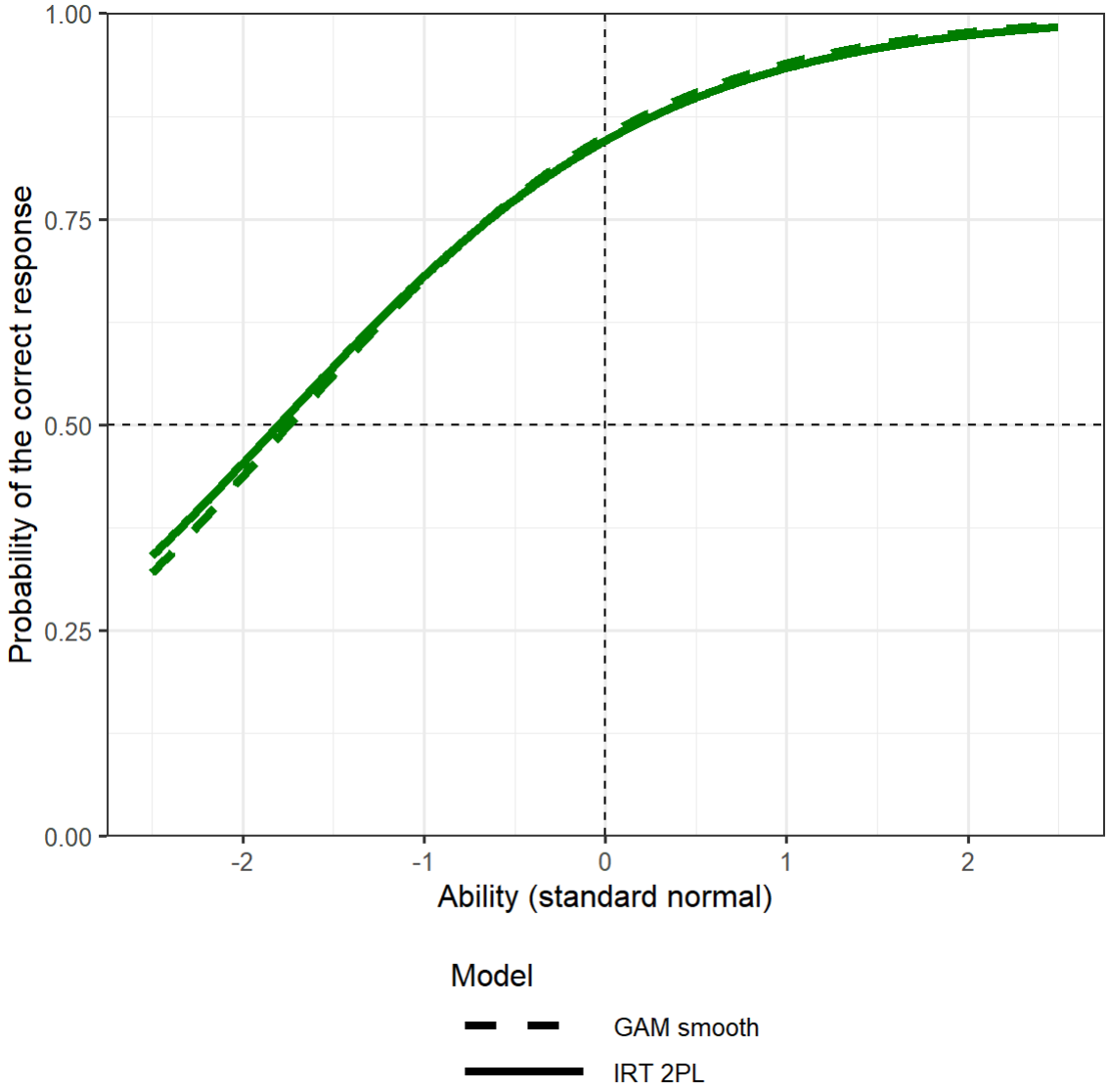
10.1.51.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	11.5	22.7	43.9	156.3	3.1	0.9
Middle quartiles	5.0	30.7	47.0	68.2	207.6	3.7	0.7
Upper quartile	4.3	41.8	56.9	80.3	286.7	4.1	0.6
Overall	2.7	25.9	44.7	67.4	286.7	3.7	0.8

10.1.52 M041177 (TIMSS)

10.1.52.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	81.54%	299	TIMSS	MC	0.35	0.30	0.94	−1.80	0.011

10.1.52.2 Gender DIF:

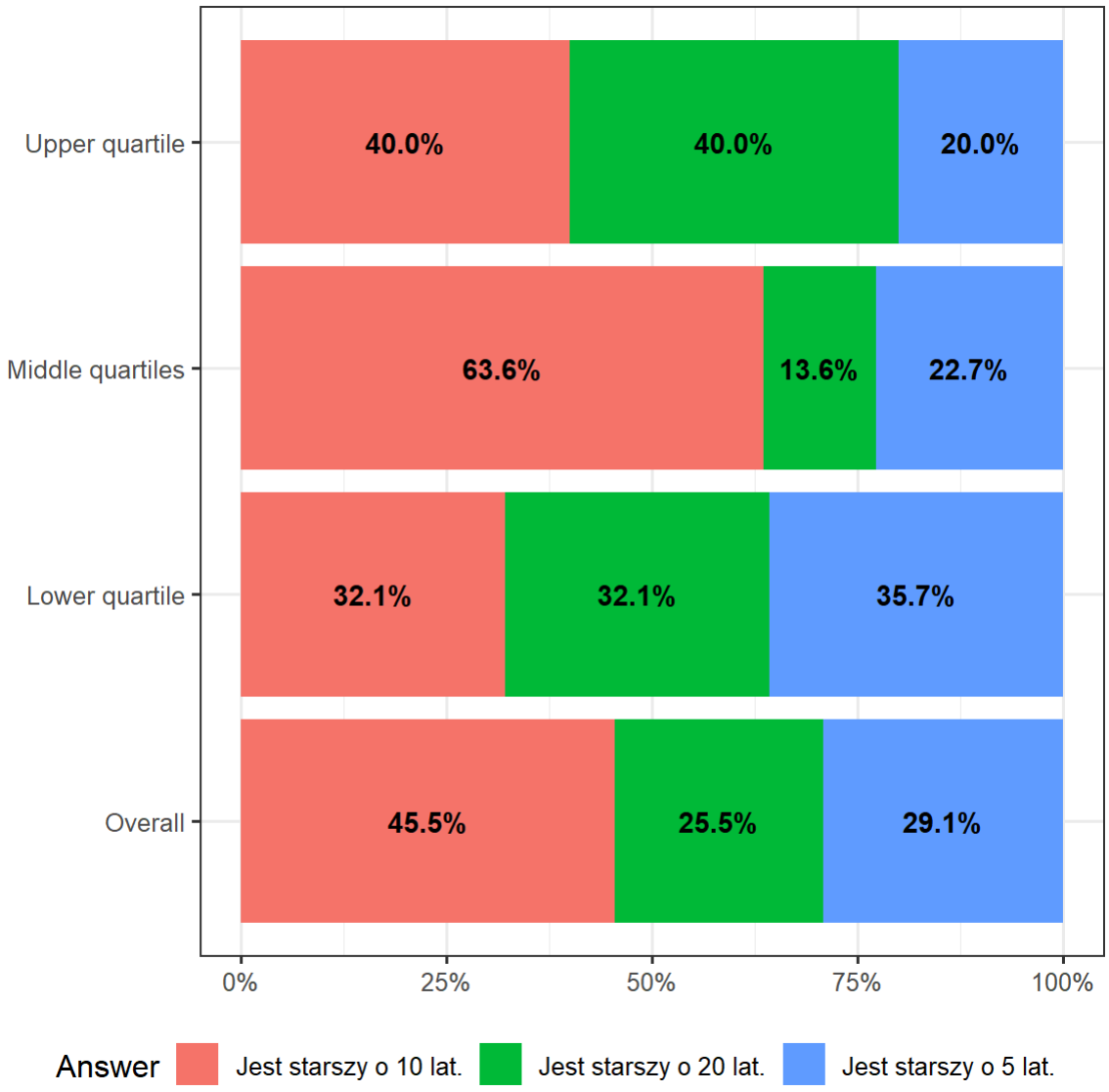
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.29	0.318	1.000	0.009	N/A

10.1.52.3 Frequency of answers:

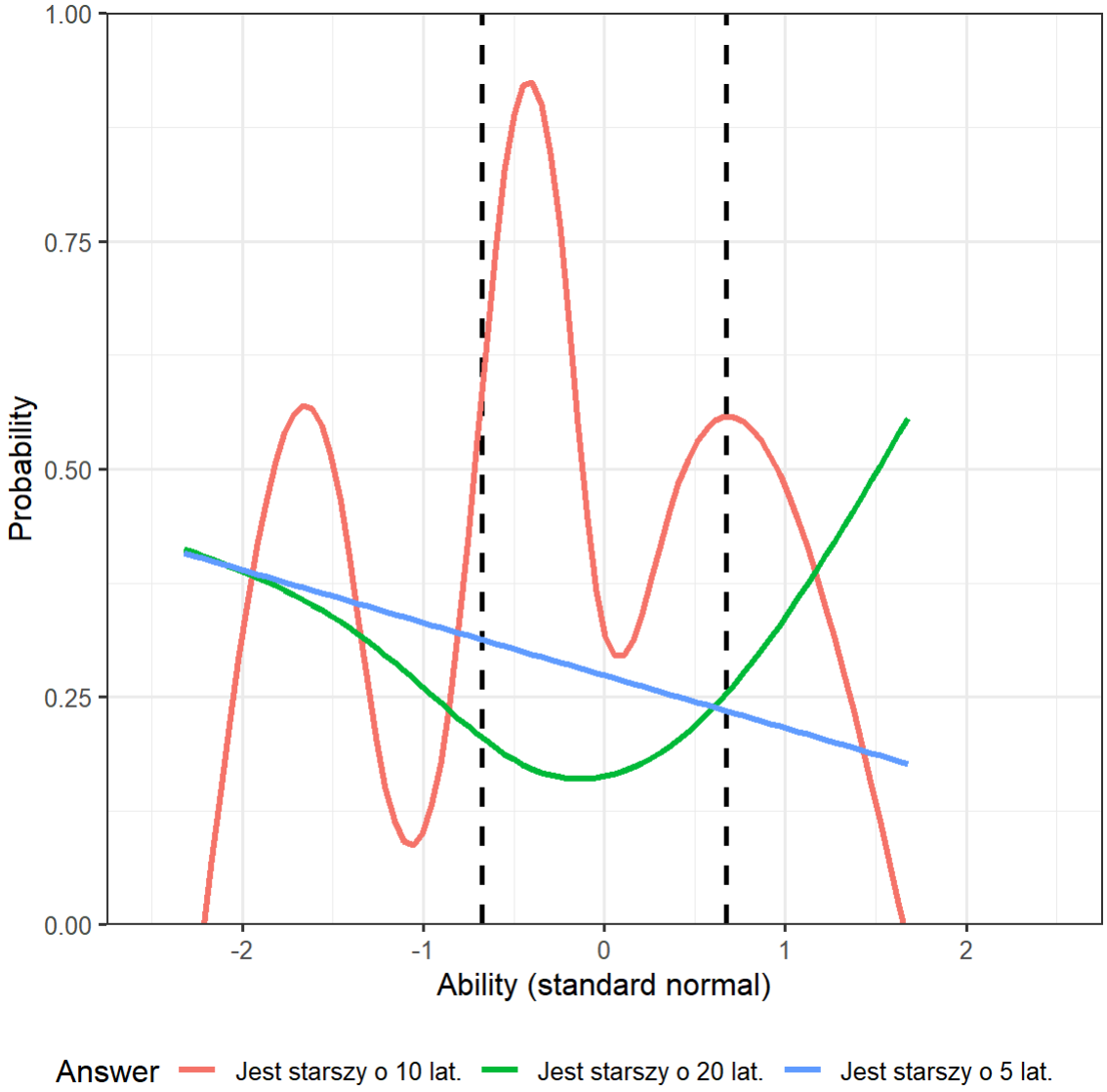
Answer	Frequency	Percent
Jest starszy o 10 lat.	25	8.4%
Jest starszy o 15 lat. (correct)	243	81.5%
Jest starszy o 20 lat.	14	4.7%
Jest starszy o 5 lat.	16	5.4%
Overall	298	100.0%

10.1.52.4 Distractors functioning:

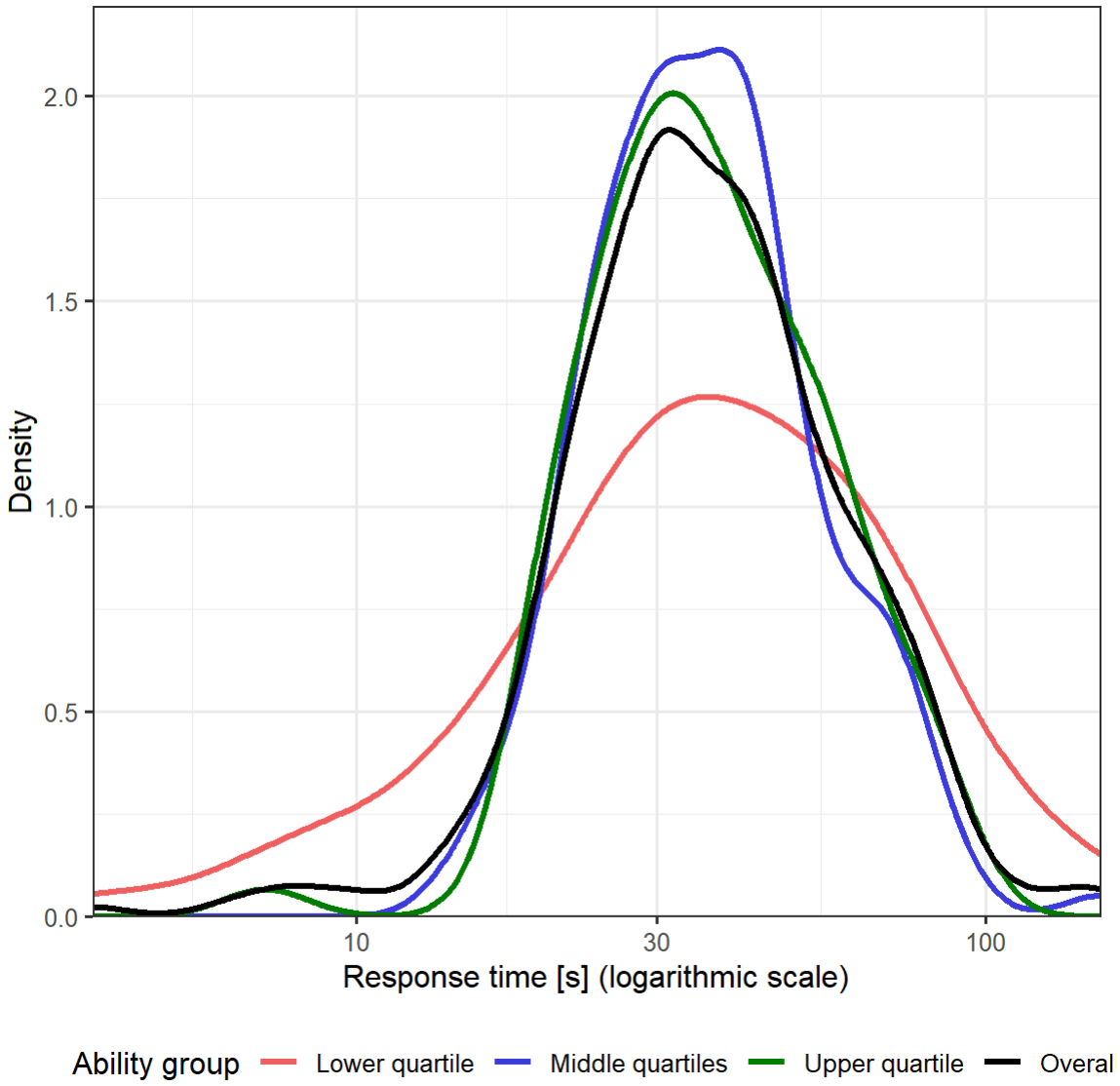
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Jest starszy o 10 lat.	9	14	2	25	32.1%	63.6%	40.0%	45.5%
Jest starszy o 20 lat.	9	3	2	14	32.1%	13.6%	40.0%	25.5%
Jest starszy o 5 lat.	10	5	1	16	35.7%	22.7%	20.0%	29.1%
Overall	28	22	5	55	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



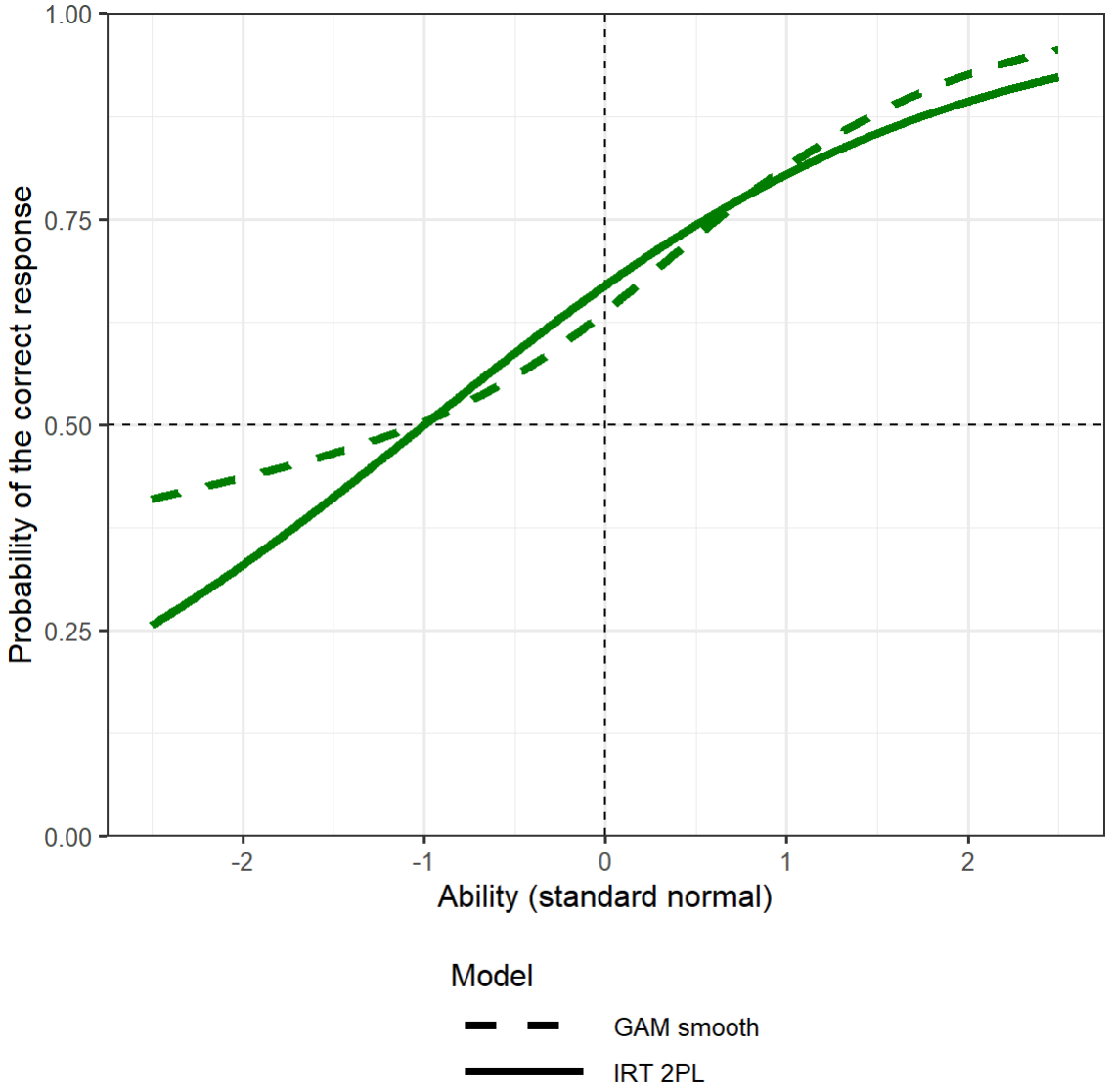
10.1.52.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.8	22.6	35.3	58.3	146.8	3.5	0.7
Middle quartiles	14.1	27.1	35.6	45.8	152.4	3.6	0.4
Upper quartile	7.2	26.8	35.0	50.9	86.7	3.6	0.4
Overall	3.8	26.2	35.2	48.4	152.4	3.6	0.5

10.1.53 M100039 (LLM)

10.1.53.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	65.52%	319	LLM simple setup	MC	0.29	0.26	0.71	−1.00	0.010

10.1.53.2 Gender DIF:

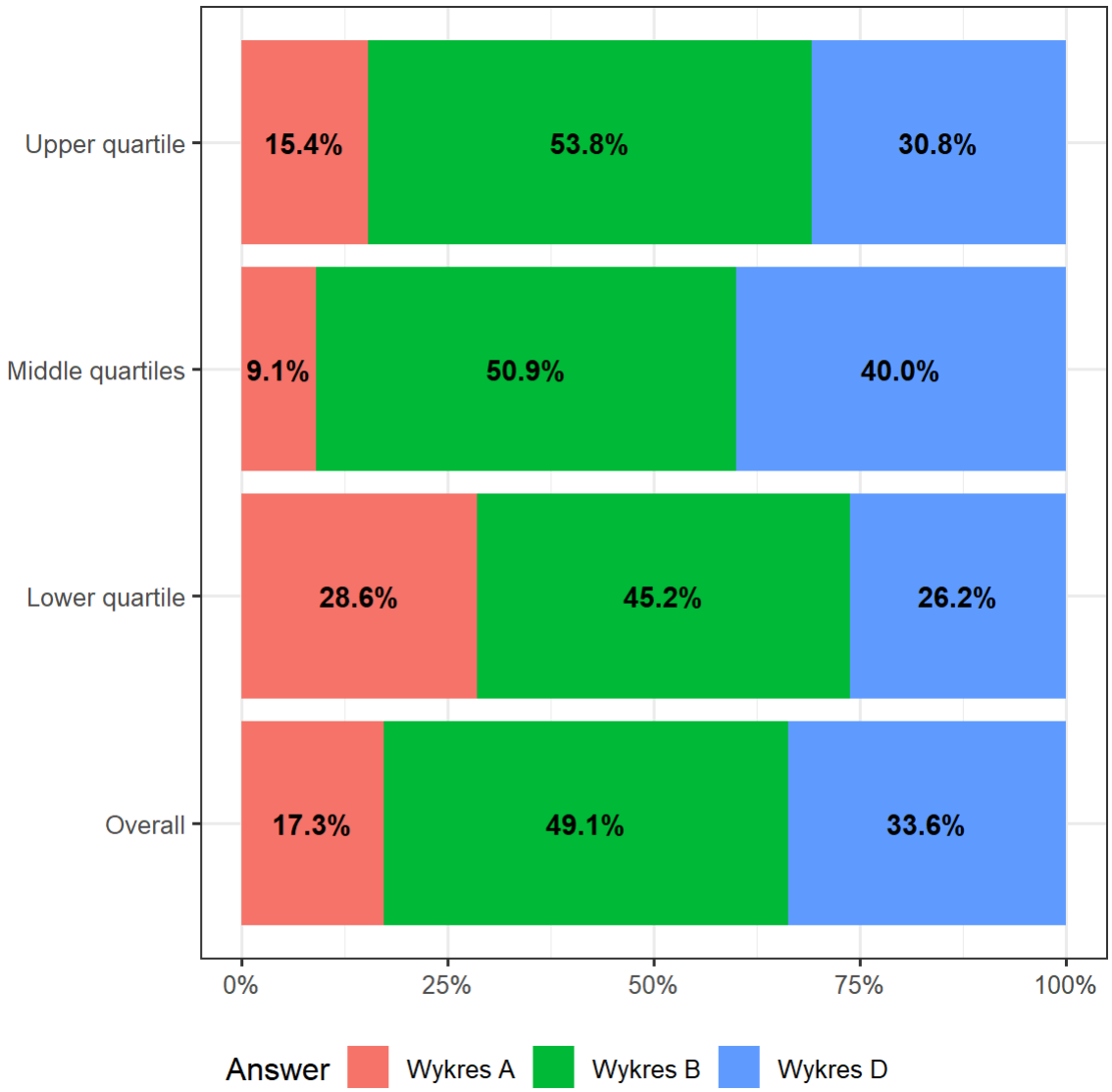
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.44	0.802	1.000	0.001	N/A

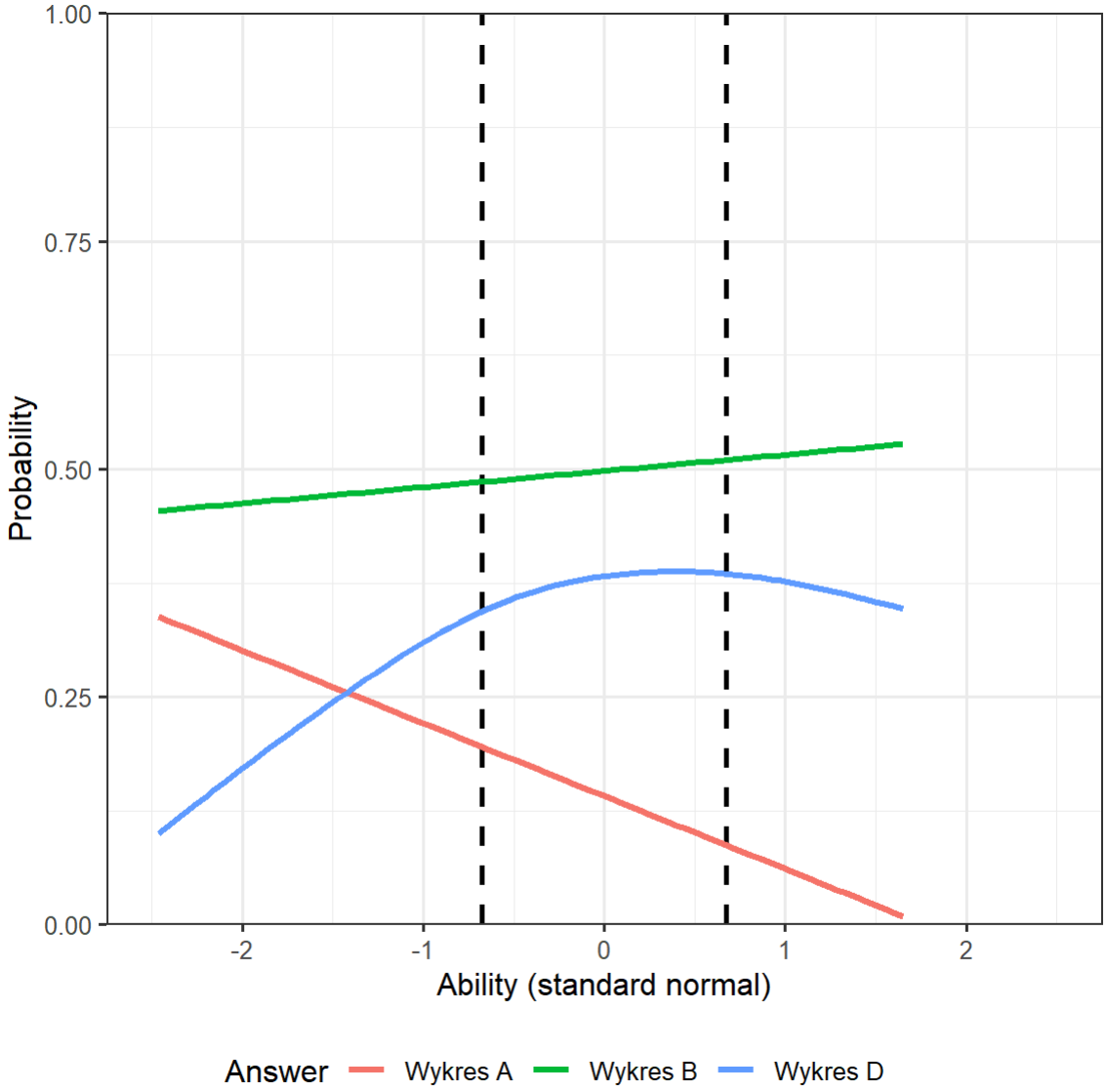
10.1.53.3 Frequency of answers:

Answer	Frequency	Percent
Wykres A	19	6.0%
Wykres B	54	16.9%
Wykres C (correct)	209	65.5%
Wykres D	37	11.6%
Overall	319	100.0%

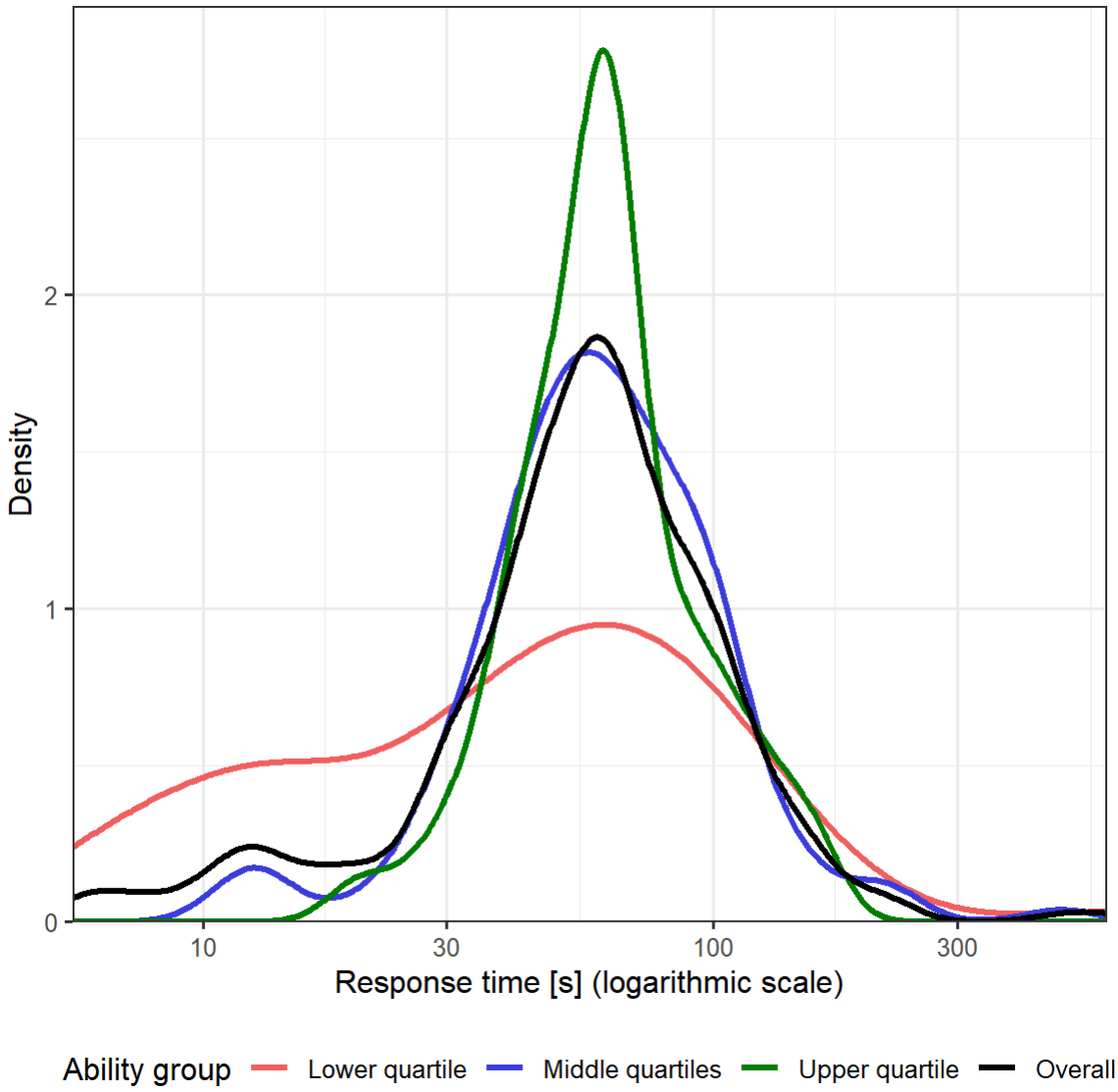
10.1.53.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Wykres A	12	5	2	19	28.6%	9.1%	15.4%	17.3%
Wykres B	19	28	7	54	45.2%	50.9%	53.8%	49.1%
Wykres D	11	22	4	37	26.2%	40.0%	30.8%	33.6%
Overall	42	55	13	110	100.0%	100.0%	100.0%	100.0%


Upper quartile
Middle quartiles
Lower quartile
Overall
15.4%
53.8%
30.8%
9.1%
50.9%
40.0%
28.6%
45.2%
26.2%
17.3%
49.1%
33.6%
0%
25%
50%
75%
100%
Answer
Wykres A
Wykres B
Wykres D



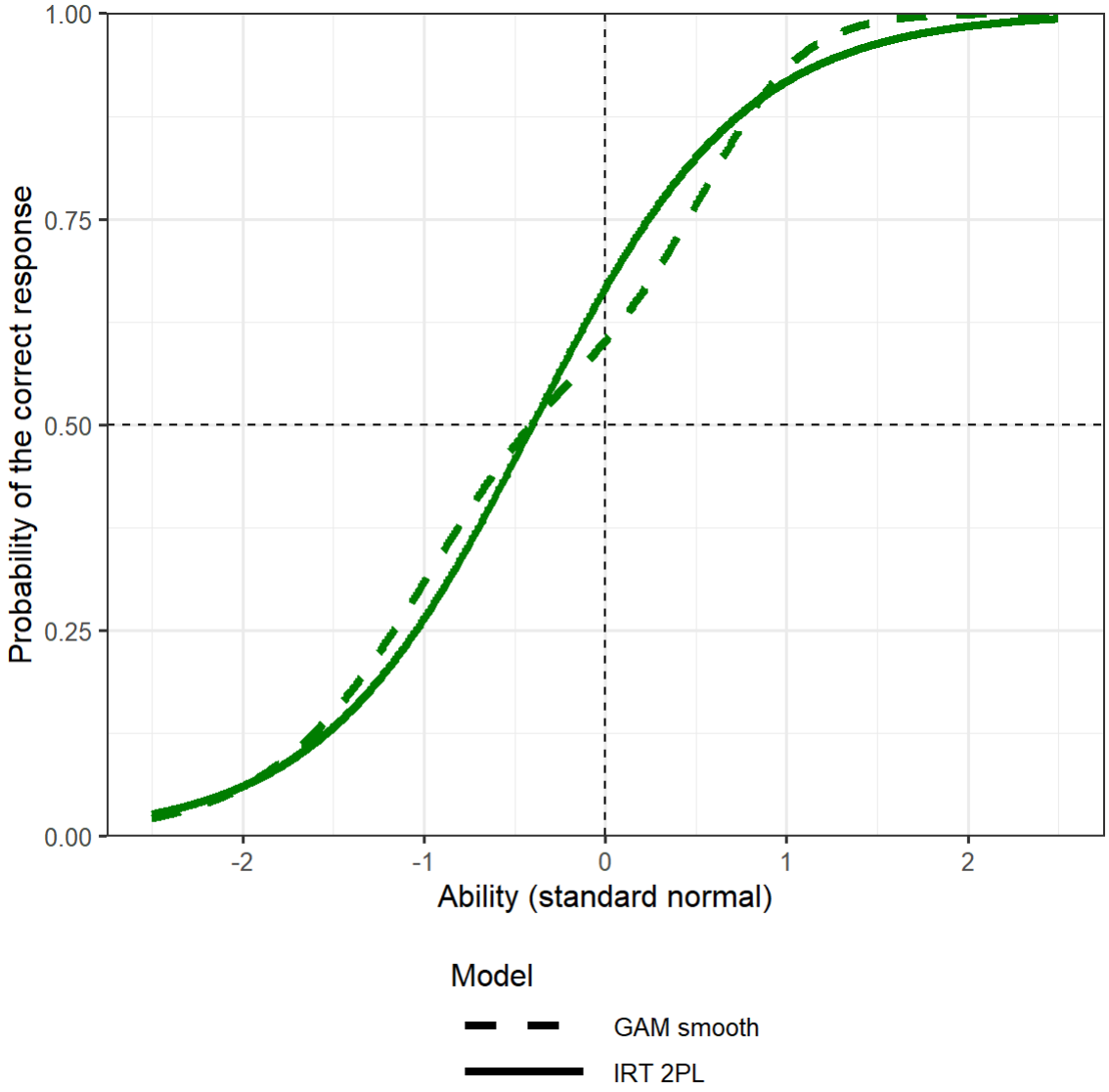
10.1.53.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.5	17.0	46.1	79.7	587.8	3.6	1.0
Middle quartiles	11.0	43.5	58.3	83.9	481.2	4.1	0.6
Upper quartile	19.4	46.1	60.8	75.8	160.5	4.1	0.4
Overall	5.5	40.8	57.9	81.7	587.8	4.0	0.7

10.1.54 M041271 (TIMSS)

10.1.54.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	60.53%	304	TIMSS	MC	0.54	0.56	1.72	−0.40	0.000

10.1.54.2 Gender DIF:

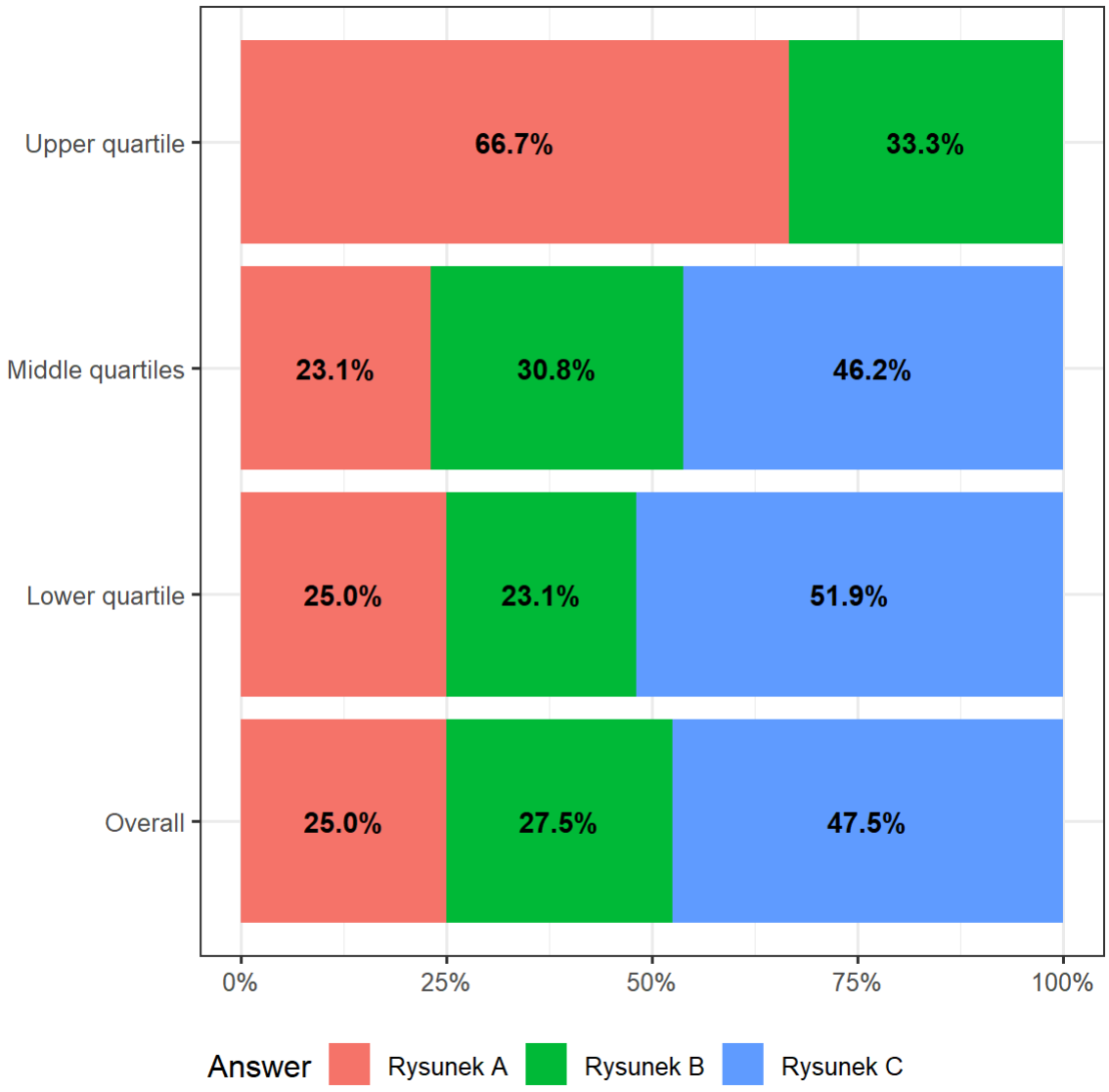
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
8.59	0.014	1.000	0.022	N/A

10.1.54.3 Frequency of answers:

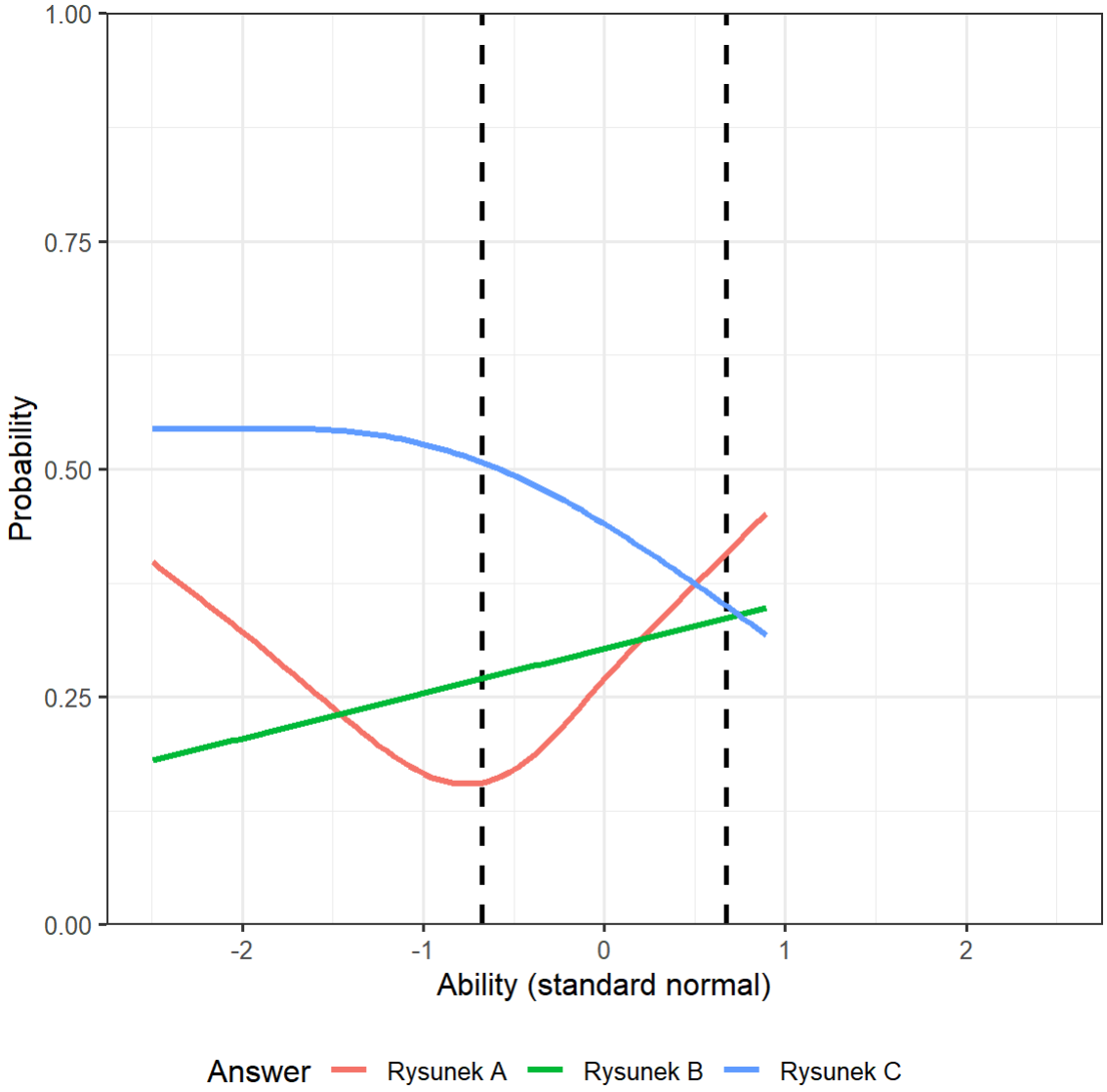
Answer	Frequency	Percent
Rysunek A	30	9.9%
Rysunek B	33	10.9%
Rysunek C	57	18.8%
Rysunek D (correct)	184	60.5%
Overall	304	100.0%

10.1.54.4 Distractors functioning:

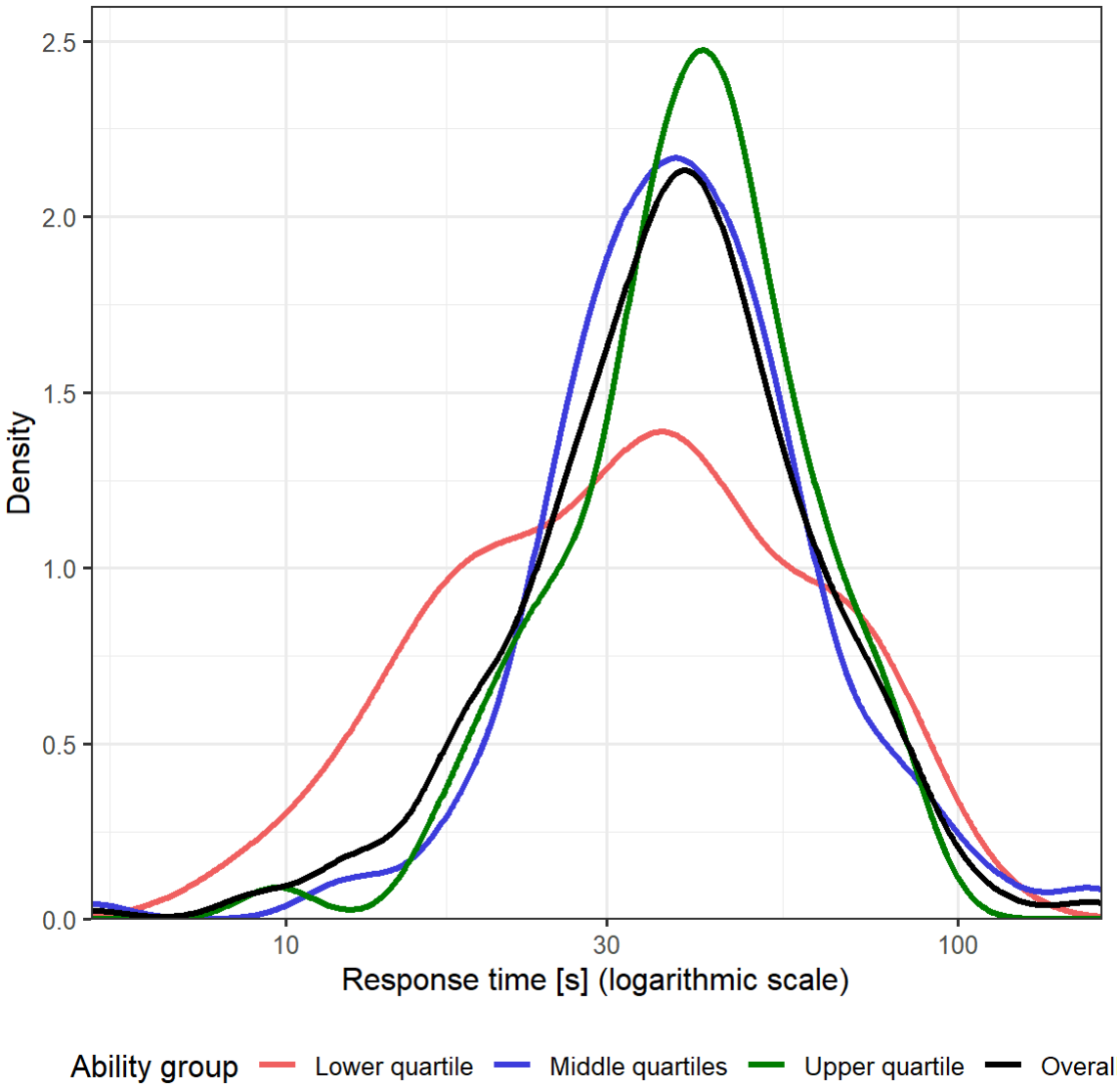
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Rysunek A	13	15	2	30	25.0%	23.1%	66.7%	25.0%
Rysunek B	12	20	1	33	23.1%	30.8%	33.3%	27.5%
Rysunek C	27	30	0	57	51.9%	46.2%	0.0%	47.5%
Overall	52	65	3	120	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



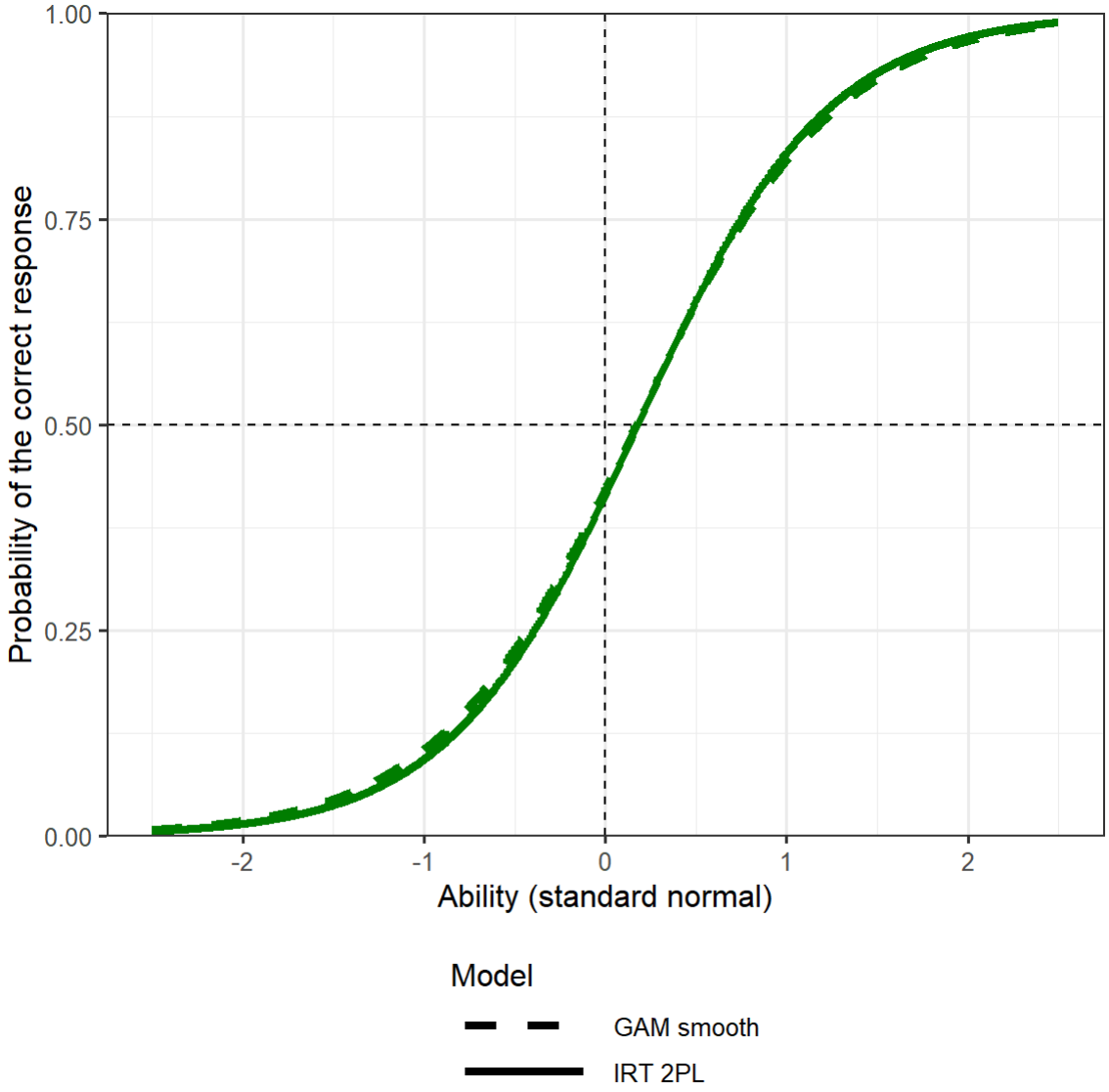
10.1.54.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	8.6	19.8	33.3	48.6	97.0	3.5	0.6
Middle quartiles	5.1	29.3	38.3	50.5	163.2	3.7	0.5
Upper quartile	9.7	32.2	40.3	51.4	86.3	3.7	0.4
Overall	5.1	28.2	38.0	50.8	163.2	3.6	0.5

10.1.55 M061031 (TIMSS)

10.1.55.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	46.26%	294	TIMSS	MC	0.59	0.68	1.93	0.18	0.027

10.1.55.2 Gender DIF:

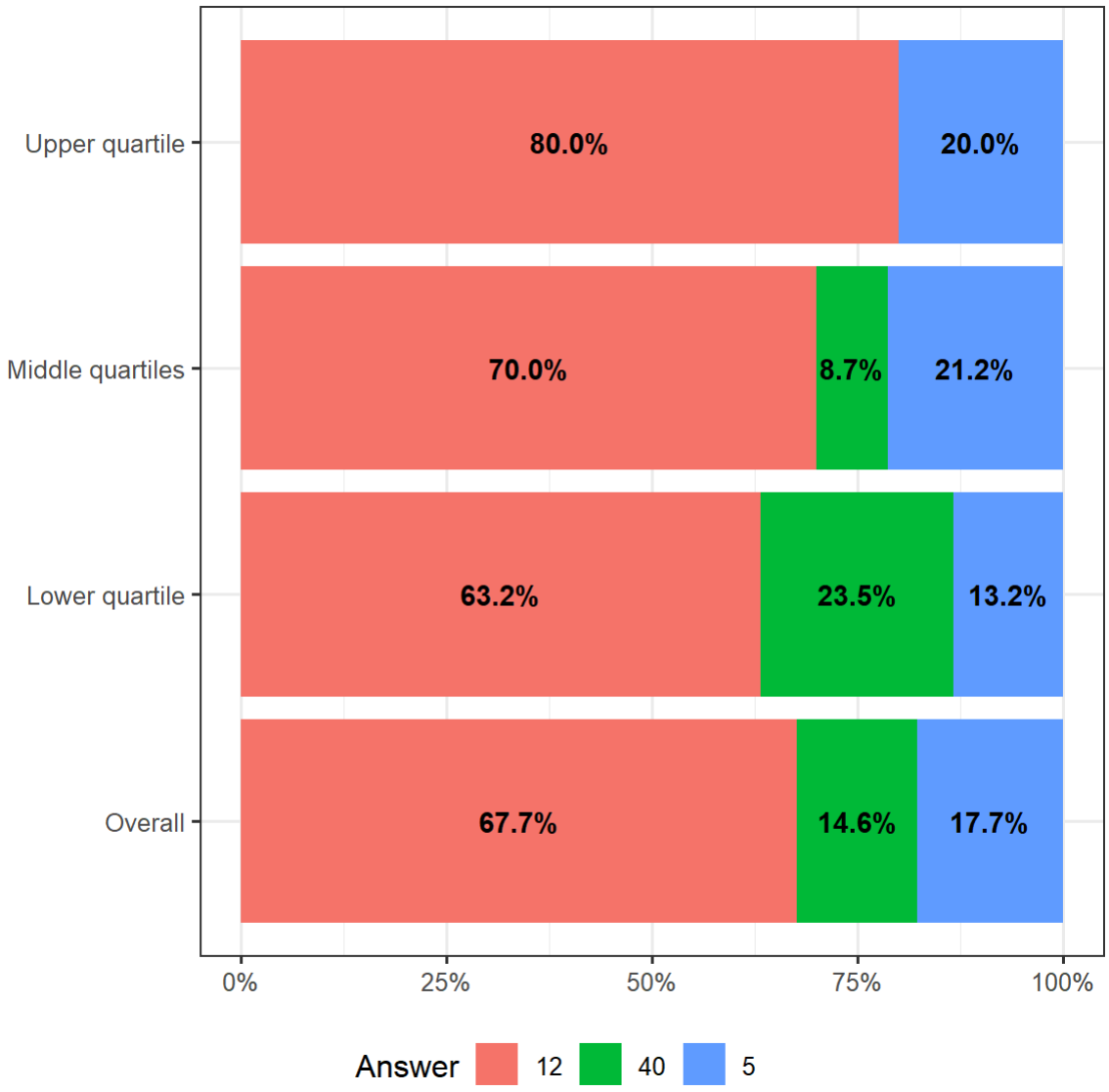
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.24	0.888	1.000	0.001	N/A

10.1.55.3 Frequency of answers:

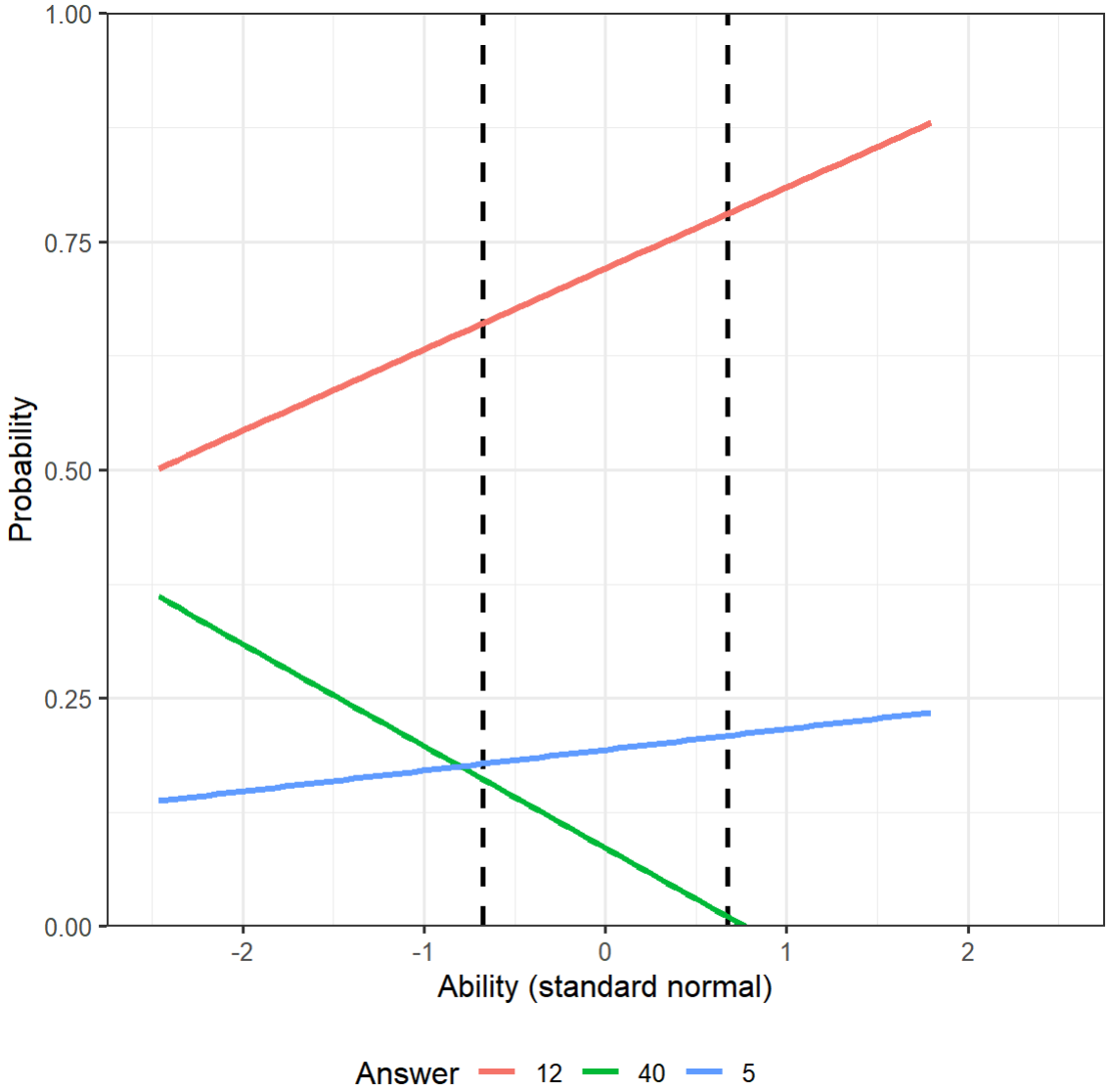
Answer	Frequency	Percent
4 (correct)	136	46.3%
5	28	9.5%
12	107	36.4%
40	23	7.8%
Overall	294	100.0%

10.1.55.4 Distractors functioning:

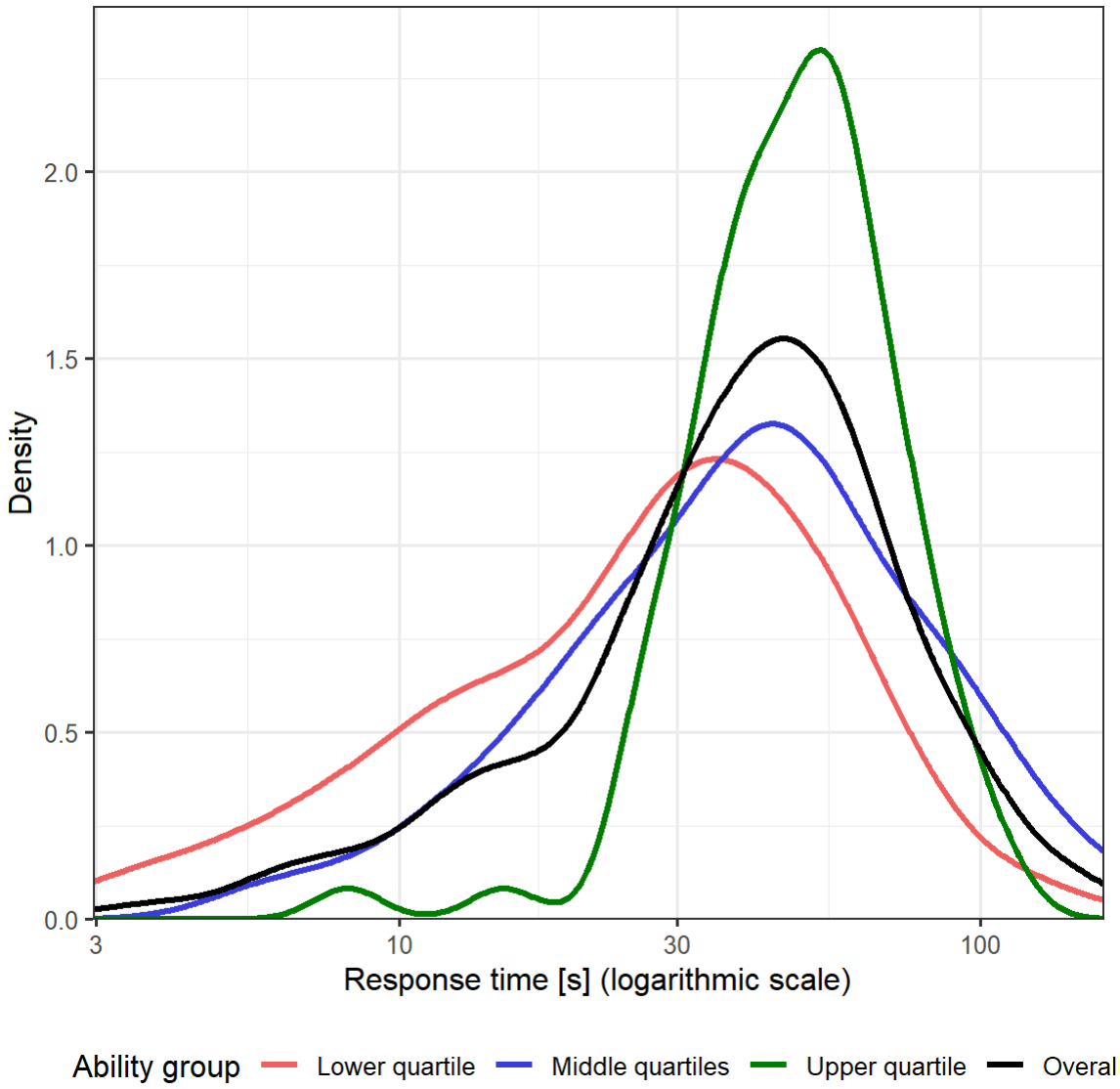
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
12	43	56	8	107	63.2%	70.0%	80.0%	67.7%
40	16	7	0	23	23.5%	8.8%	0.0%	14.6%
5	9	17	2	28	13.2%	21.2%	20.0%	17.7%
Overall	68	80	10	158	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 15 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



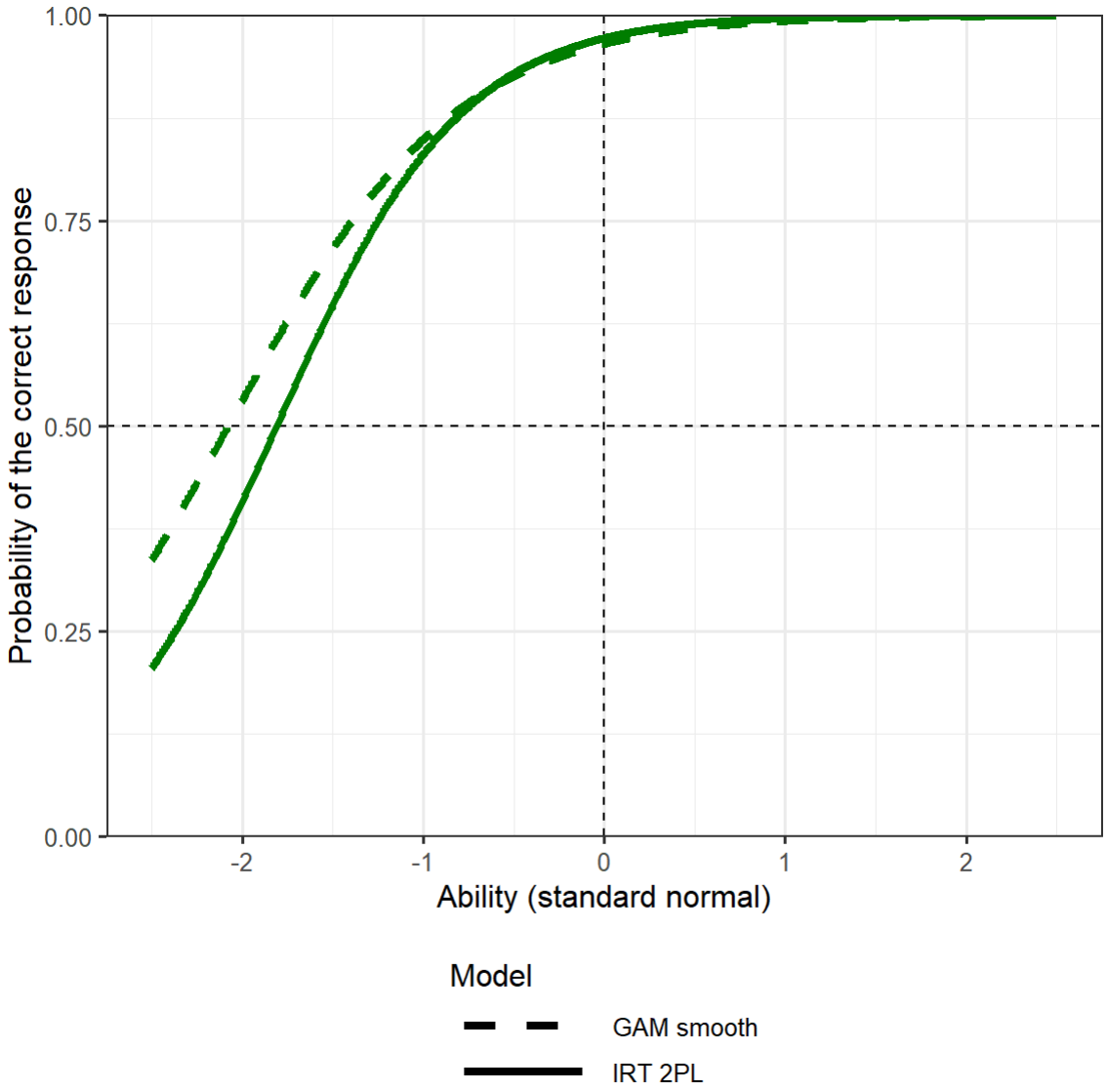
10.1.55.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	14.8	29.1	43.1	120.8	3.2	0.8
Middle quartiles	5.4	24.4	40.5	62.8	162.9	3.6	0.7
Upper quartile	8.2	36.7	48.7	60.1	110.9	3.9	0.4
Overall	3.0	25.7	40.6	57.5	162.9	3.6	0.7

10.1.56 M890024 (LLM)

10.1.56.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	91.07%	281	LLM complex setup	MC	0.41	0.31	1.96	−1.81	0.00

10.1.56.2 Gender DIF:

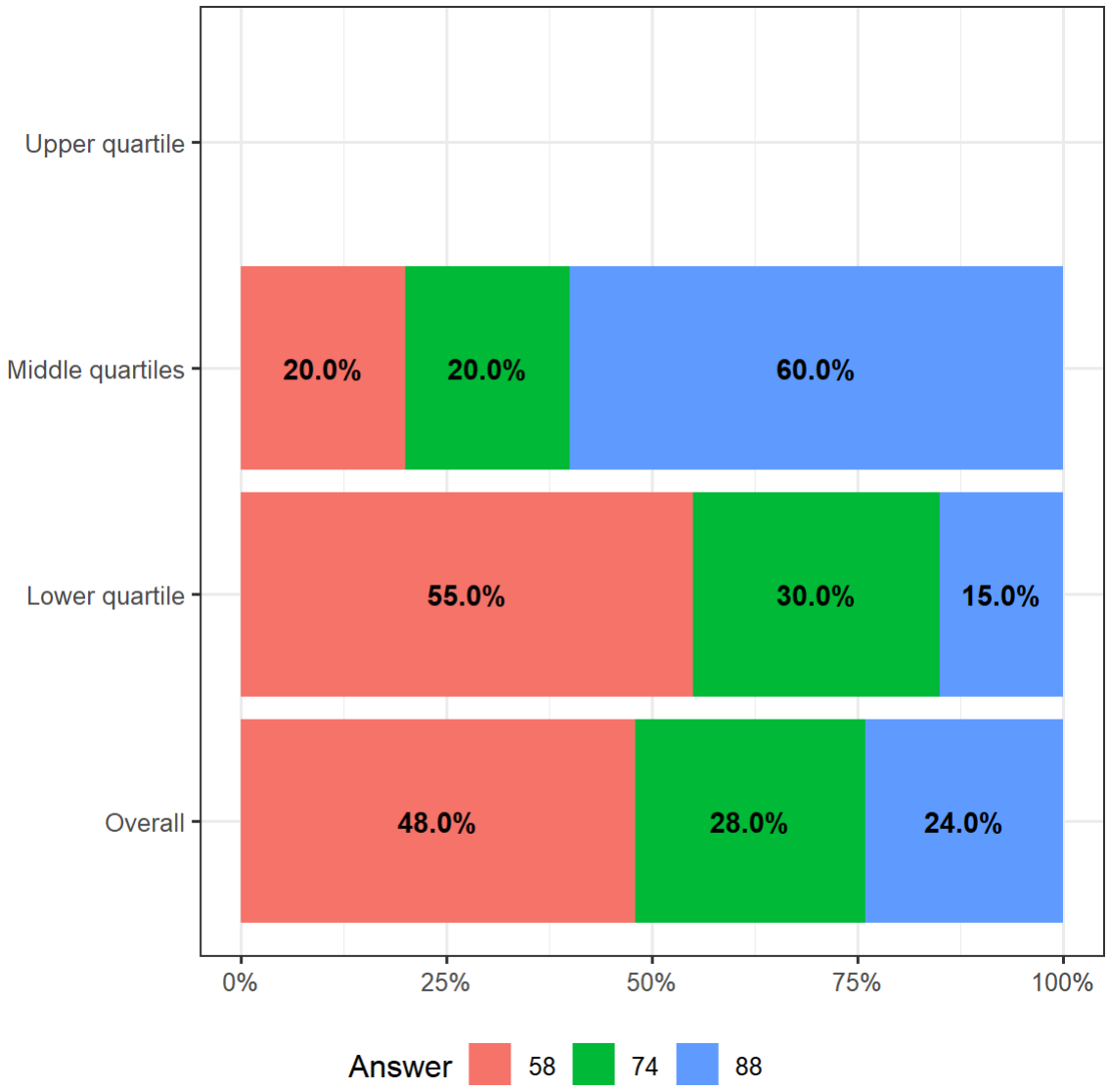
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.69	0.429	1.000	0.010	N/A

10.1.56.3 Frequency of answers:

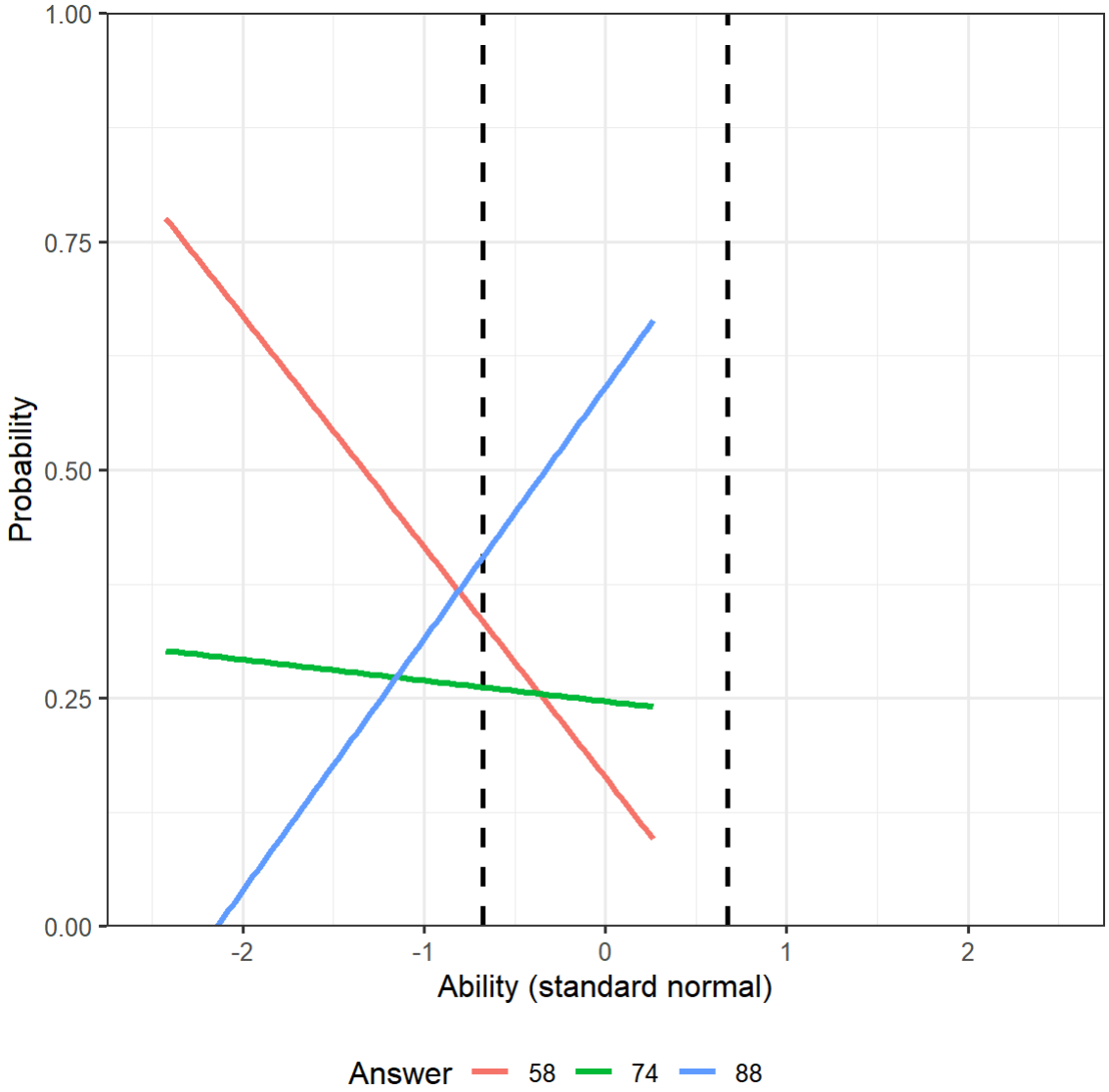
Answer	Frequency	Percent
58	12	4.3%
68 (correct)	255	91.1%
74	7	2.5%
88	6	2.1%
Overall	280	100.0%

10.1.56.4 Distractors functioning:

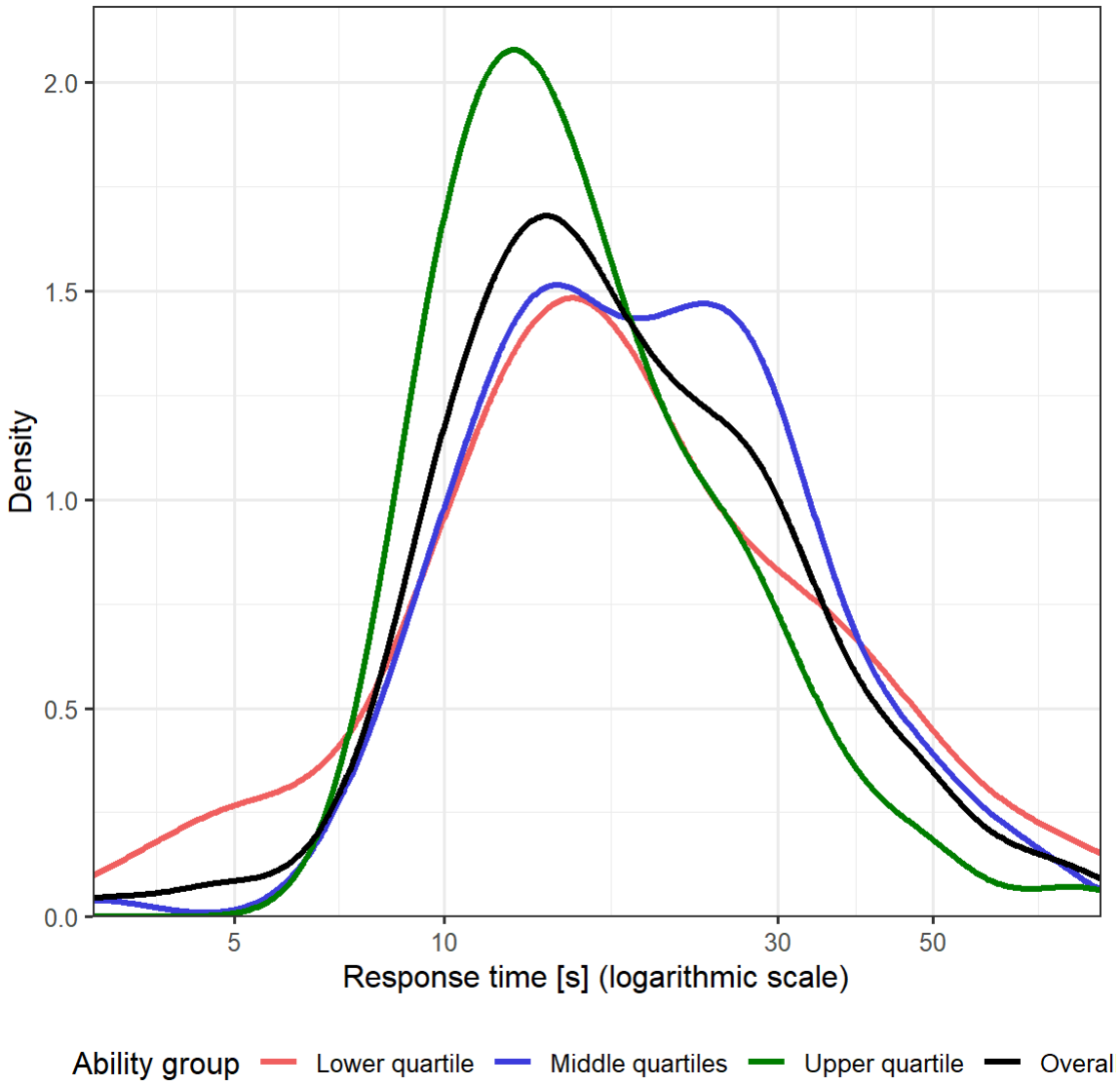
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
58	11	1	0	12	55.0%	20.0%	0.0%	48.0%
74	6	1	0	7	30.0%	20.0%	0.0%	28.0%
88	3	3	0	6	15.0%	60.0%	0.0%	24.0%
Overall	20	5	0	25	100.0%	100.0%	0.0%	100.0%




```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



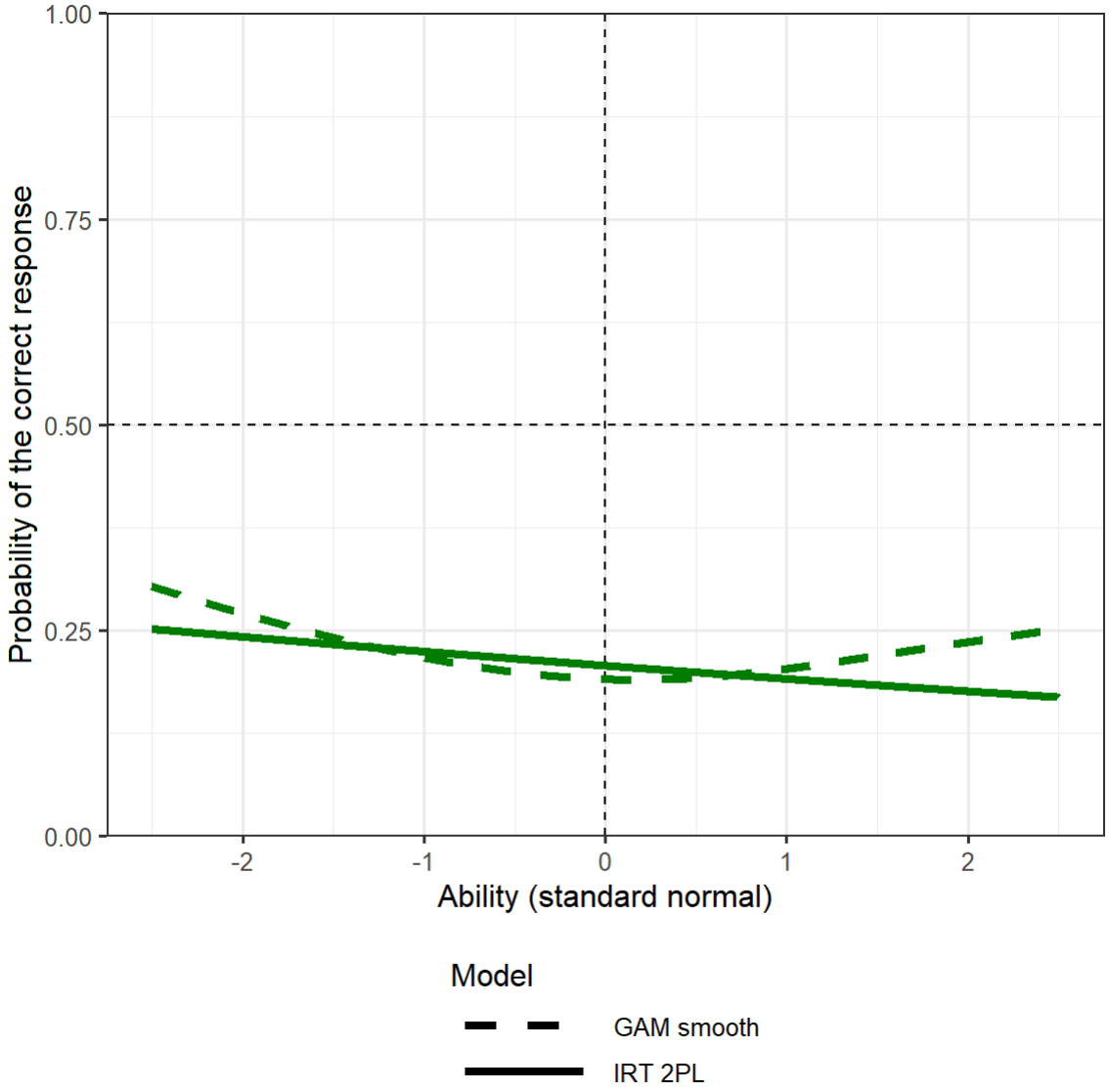
10.1.56.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	12.0	16.8	28.3	86.8	2.9	0.7
Middle quartiles	3.2	13.4	19.7	27.5	72.5	3.0	0.5
Upper quartile	7.5	11.2	14.6	21.5	80.8	2.8	0.5
Overall	3.1	12.3	16.7	26.8	86.8	2.9	0.6

10.1.57 M100009 (LLM)

10.1.57.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	20.72%	304	LLM simple setup	MC	−0.02	−0.01	−0.10	−13.31	0.018

10.1.57.2 Gender DIF:

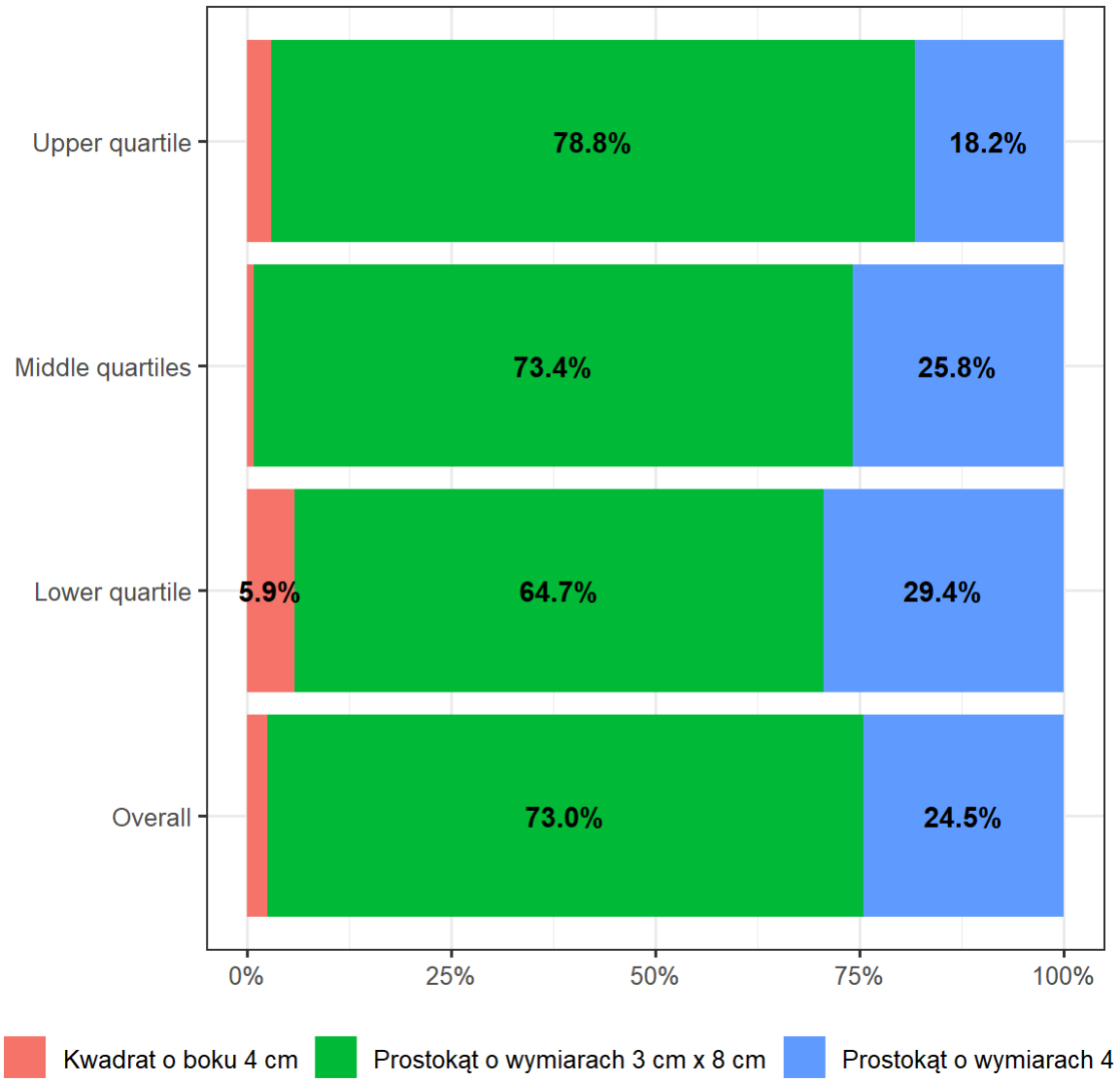
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.98	0.225	1.000	0.011	N/A

10.1.57.3 Frequency of answers:

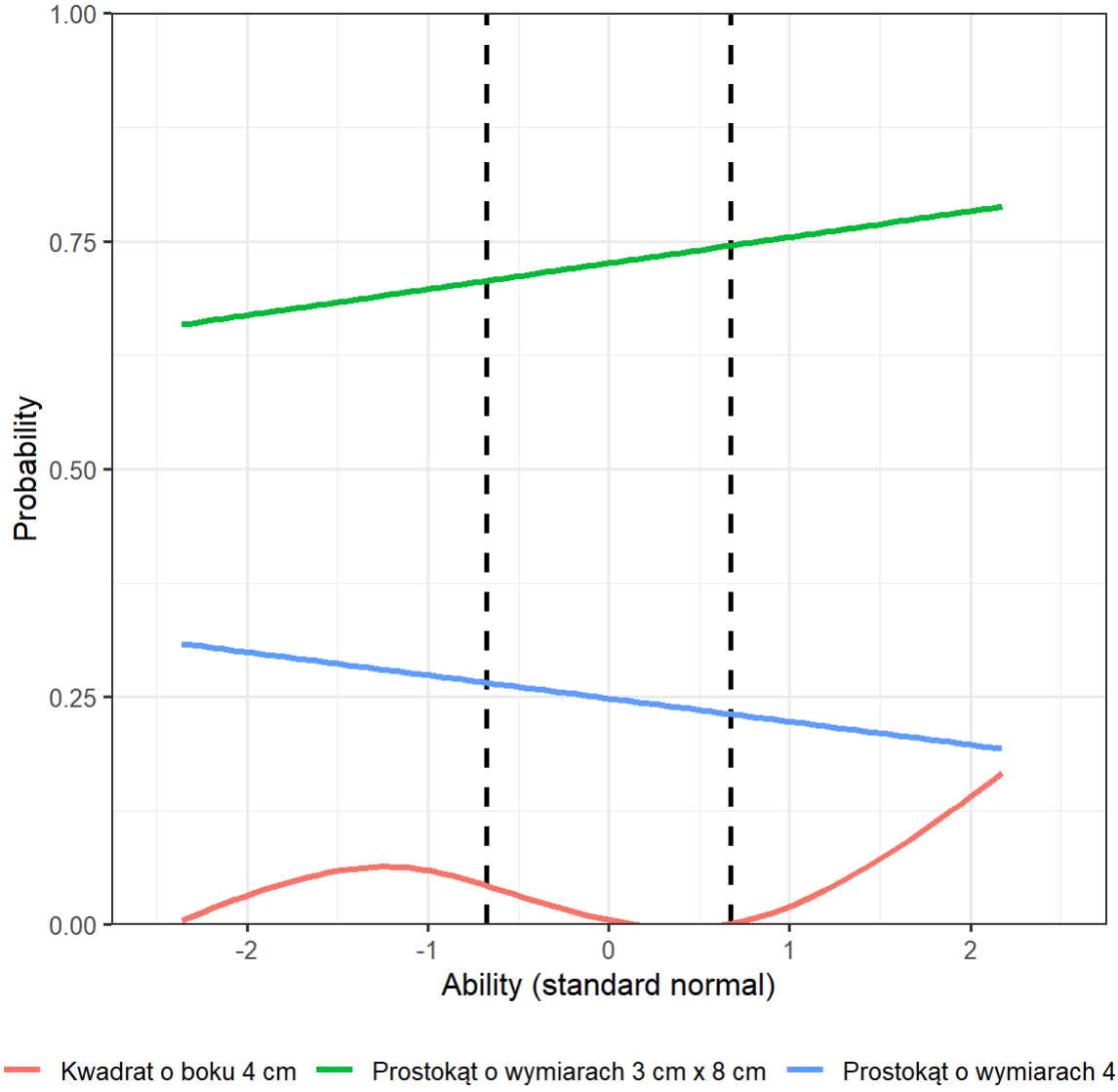
Answer	Frequency	Percent
Kwadrat o boku 4 cm	6	2.0%
Kwadrat o boku 5 cm (correct)	63	20.7%
Prostokąt o wymiarach 3 cm x 8 cm	176	57.9%
Prostokąt o wymiarach 4 cm x 6 cm	59	19.4%
Overall	304	100.0%

10.1.57.4 Distractors functioning:

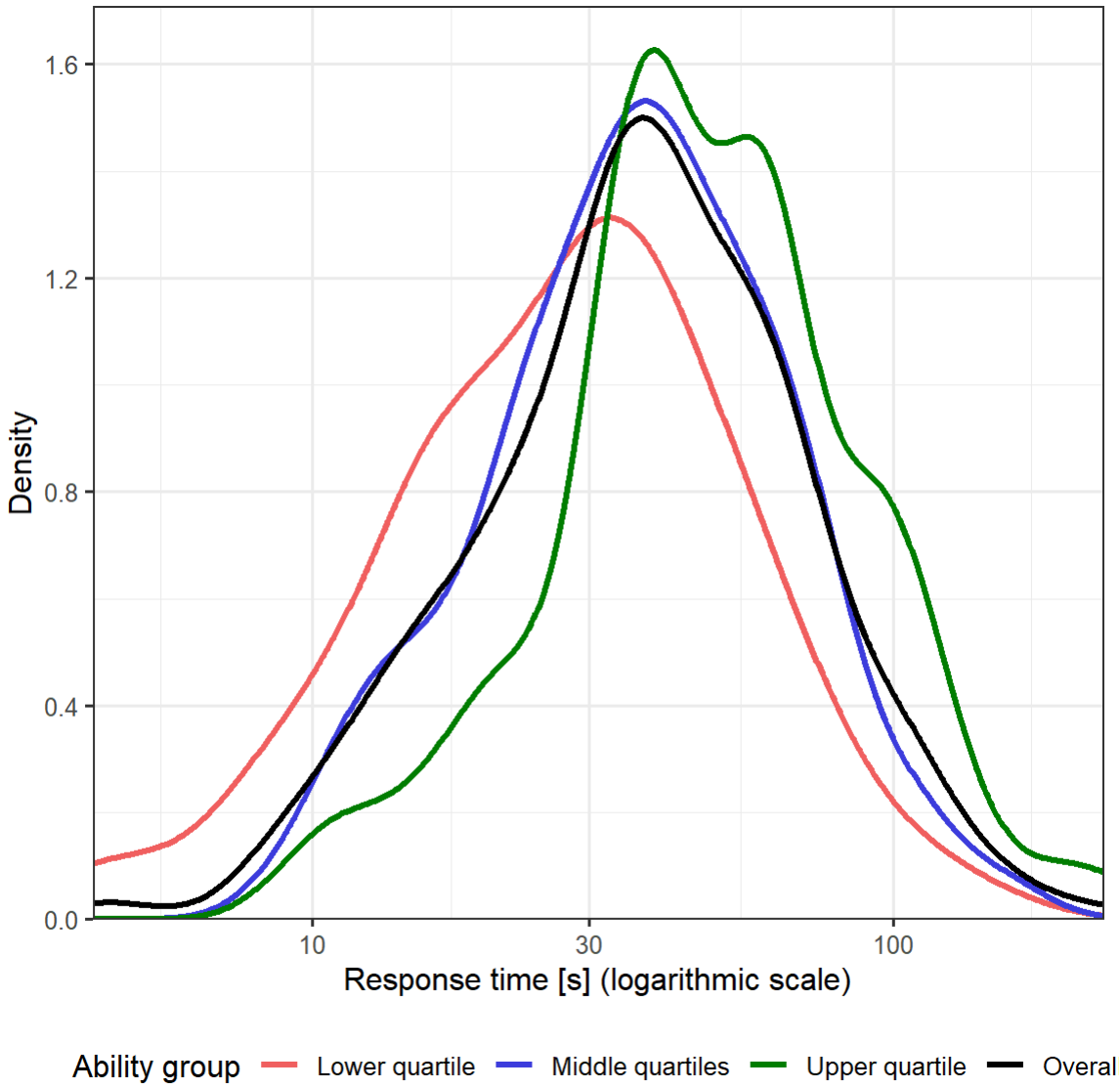
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Kwadrat o boku 4 cm	3	1	2	6	5.9%	0.8%	3.0%	2.5%
Prostokąt o wymiarach 3 cm x 8 cm	33	91	52	176	64.7%	73.4%	78.8%	73.0%
Prostokąt o wymiarach 4 cm x 6 cm	15	32	12	59	29.4%	25.8%	18.2%	24.5%
Overall	51	124	66	241	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 18 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



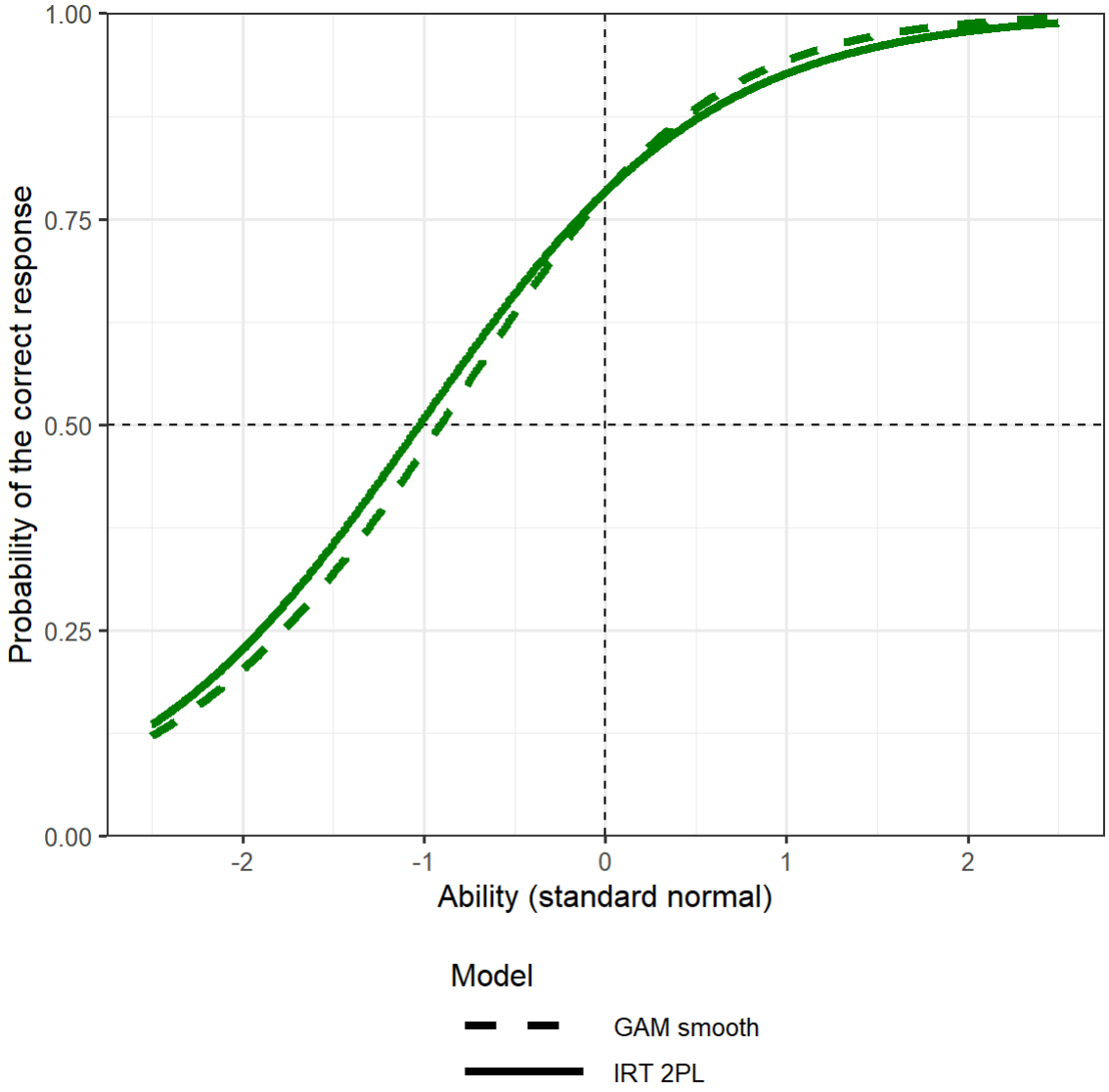
10.1.57.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.2	16.9	29.3	42.5	135.1	3.3	0.7
Middle quartiles	9.2	25.0	36.8	53.1	156.5	3.6	0.6
Upper quartile	9.3	34.2	44.9	63.8	230.2	3.8	0.6
Overall	4.2	24.1	37.1	58.3	230.2	3.6	0.7

10.1.58 M061179 (TIMSS)

10.1.58.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	72.37%	304	TIMSS	MC	0.48	0.49	1.25	−1.02	0.000

10.1.58.2 Gender DIF:

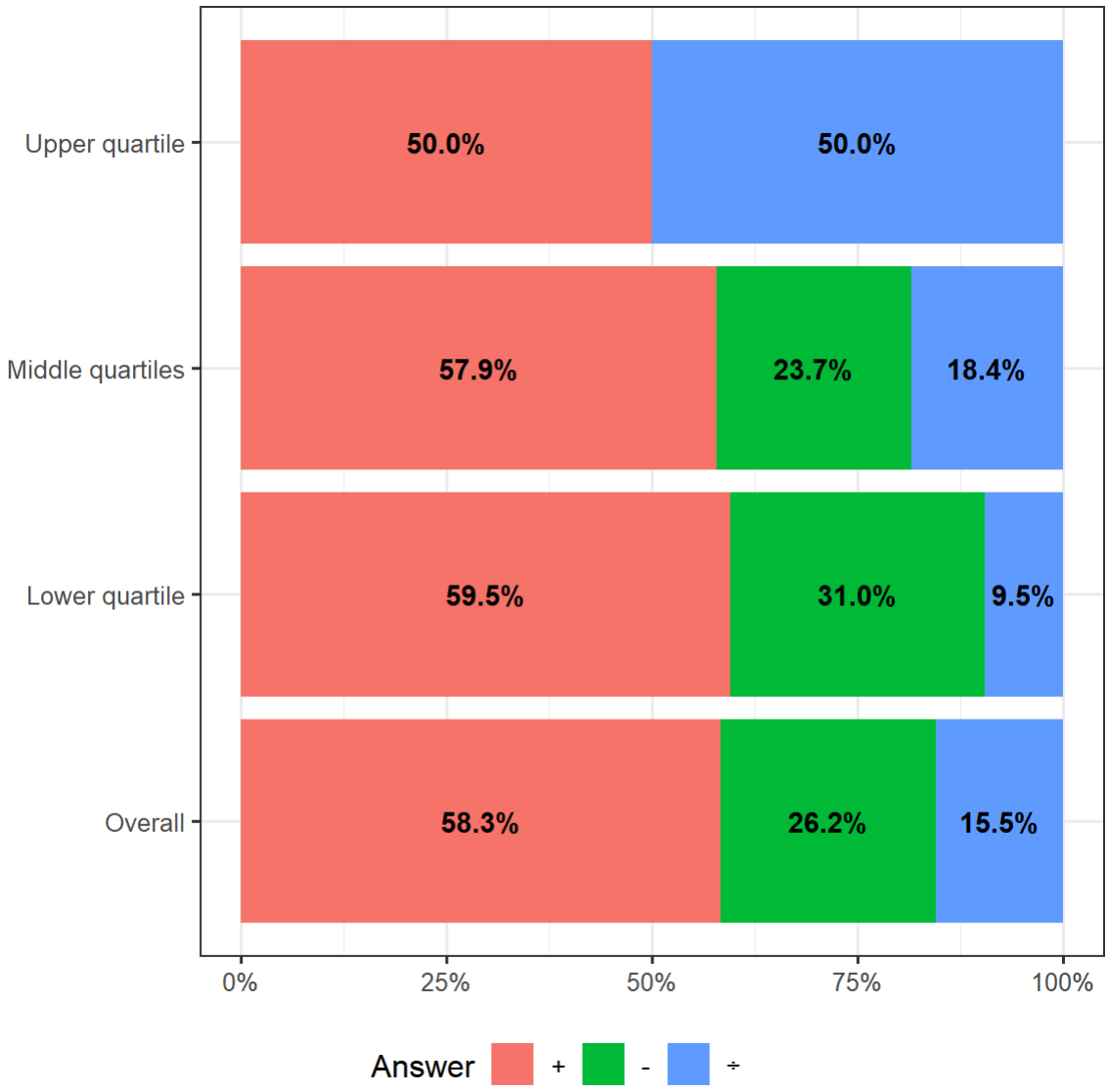
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.42	0.491	1.000	0.004	N/A

10.1.58.3 Frequency of answers:

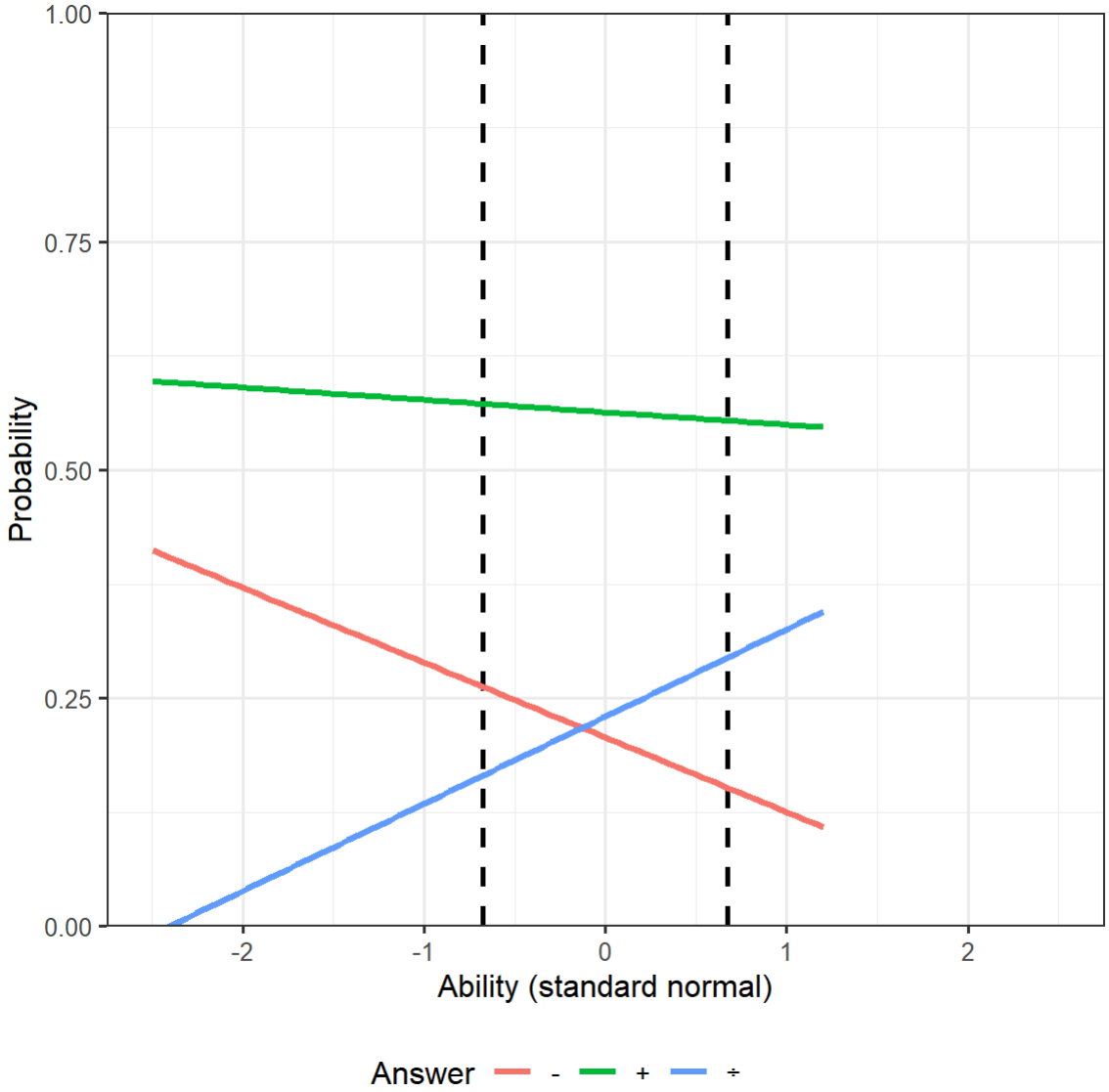
Answer	Frequency	Percent
+	49	16.1%
-	22	7.2%
× (correct)	220	72.4%
÷	13	4.3%
Overall	304	100.0%

10.1.58.4 Distractors functioning:

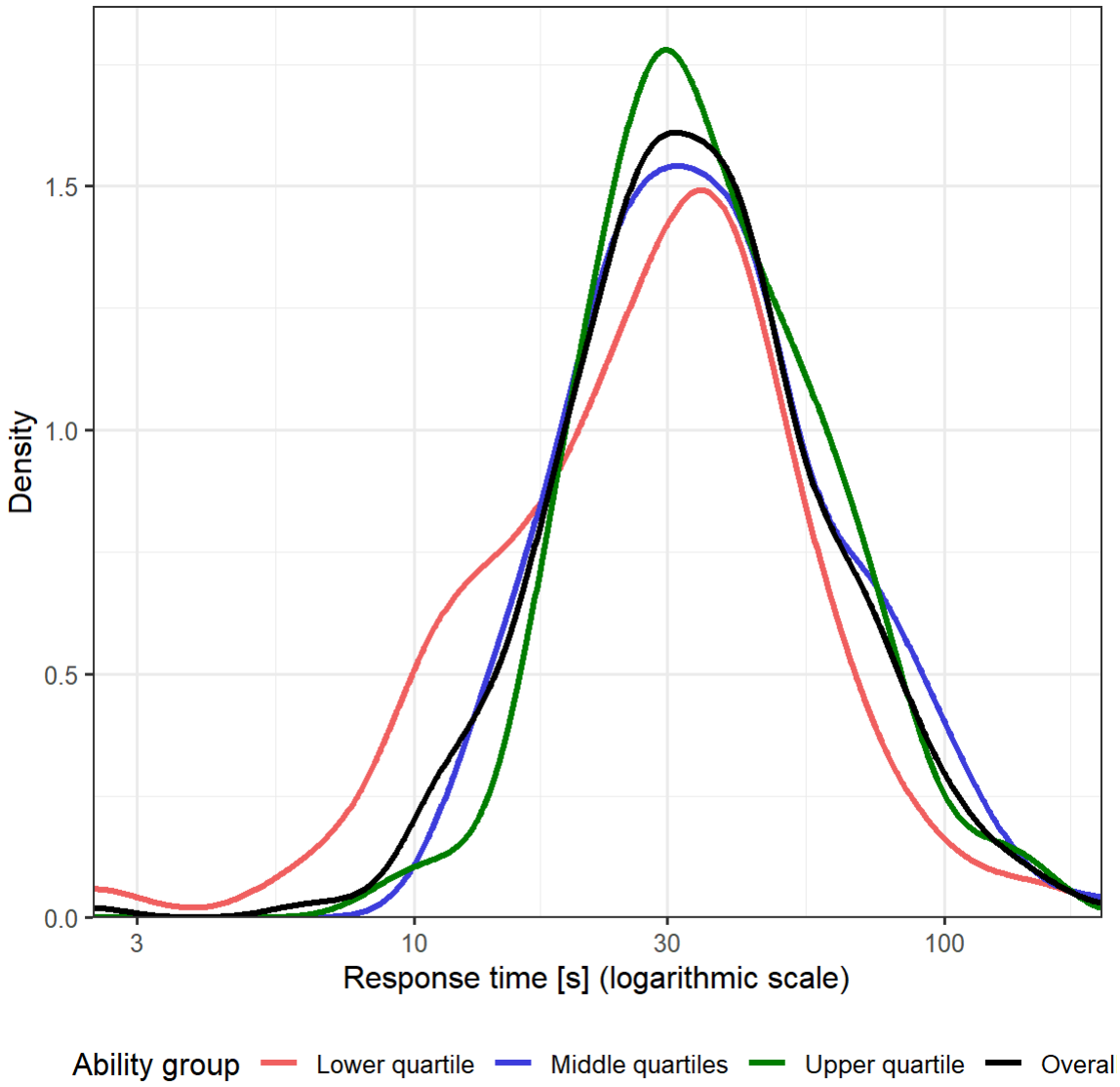
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
+	25	22	2	49	59.5%	57.9%	50.0%	58.3%
-	13	9	0	22	31.0%	23.7%	0.0%	26.2%
÷	4	7	2	13	9.5%	18.4%	50.0%	15.5%
Overall	42	38	4	84	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



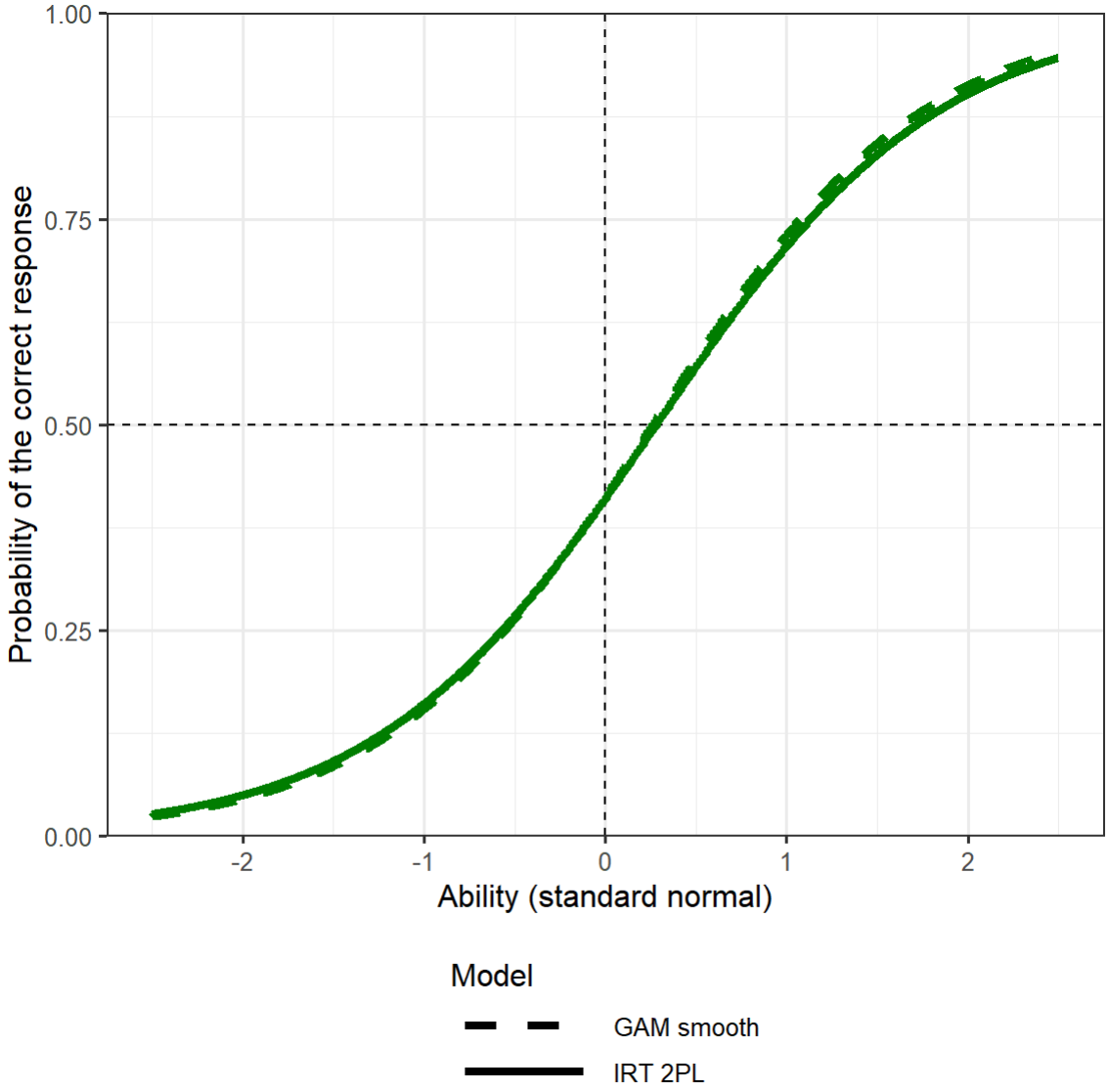
10.1.58.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.5	18.0	28.0	40.6	151.7	3.3	0.7
Middle quartiles	11.1	23.5	34.3	48.6	197.7	3.6	0.6
Upper quartile	9.6	25.8	33.5	51.0	143.6	3.6	0.5
Overall	2.5	23.3	33.6	48.1	197.7	3.5	0.6

10.1.59 M061272 (TIMSS)

10.1.59.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	44.30%	299	TIMSS	CR	0.50	0.46	1.29	0.28	0.016

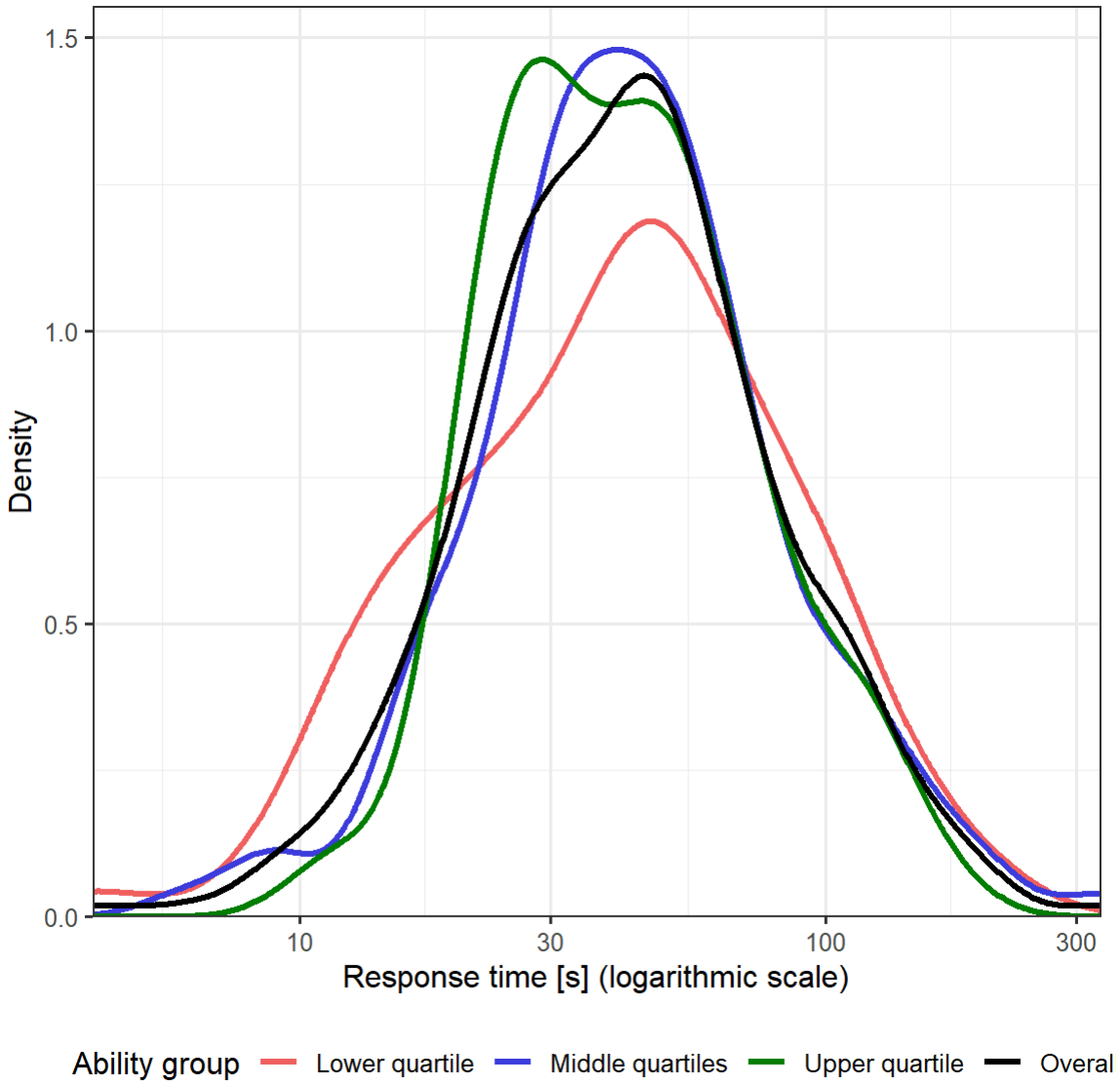
10.1.59.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.99	0.610	1.000	0.003	N/A

10.1.59.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

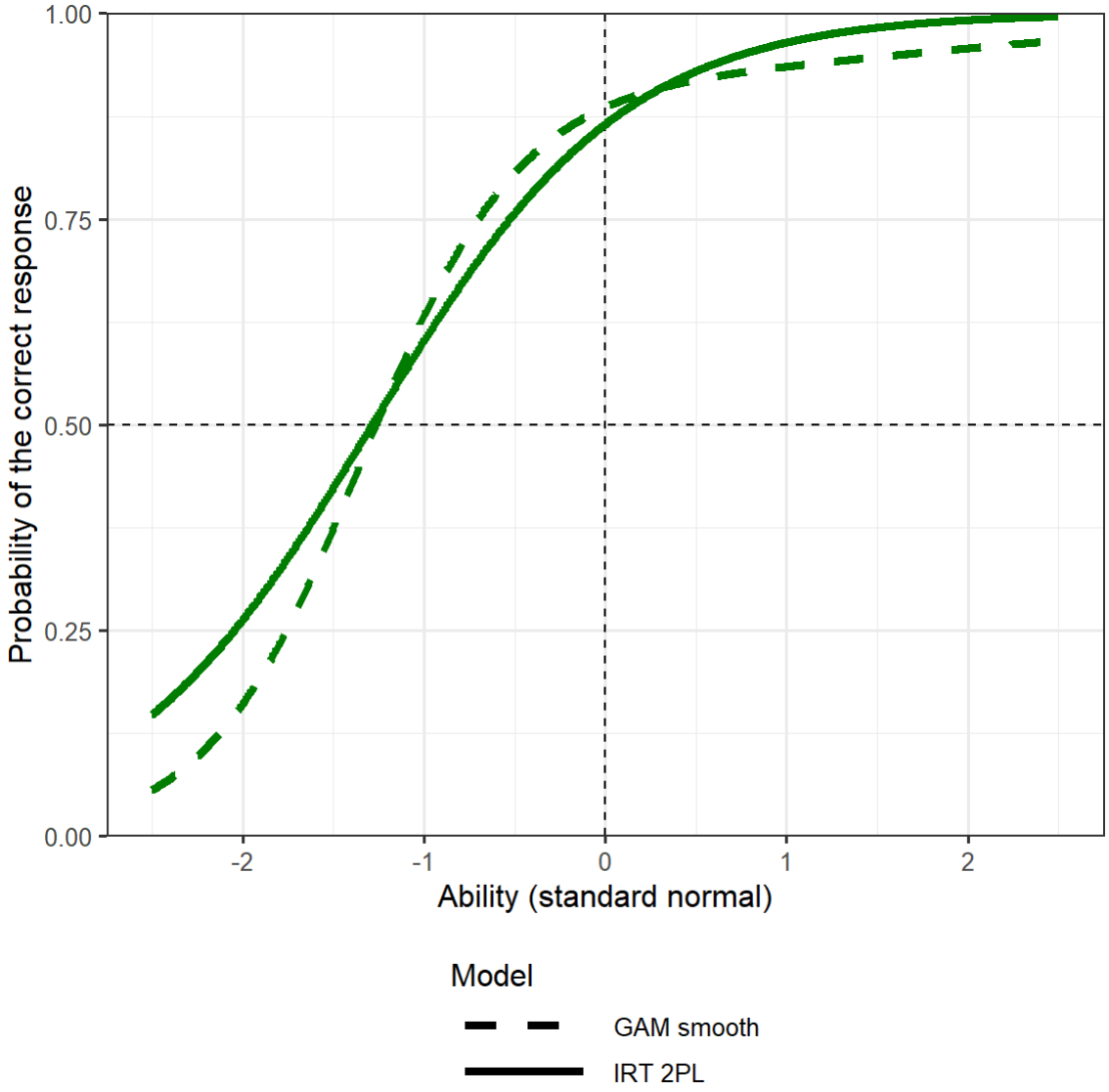
10.1.59.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.0	23.9	42.6	67.0	199.3	3.7	0.8
Middle quartiles	6.0	29.1	42.0	60.2	333.2	3.8	0.7
Upper quartile	11.1	27.3	39.9	58.0	168.5	3.7	0.6
Overall	4.0	27.1	41.8	61.0	333.2	3.7	0.7

10.1.60 M090052 (LLM)

10.1.60.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	79.19%	300	LLM teachers' selection	CR	0.45	0.33	1.45	−1.29	0.03

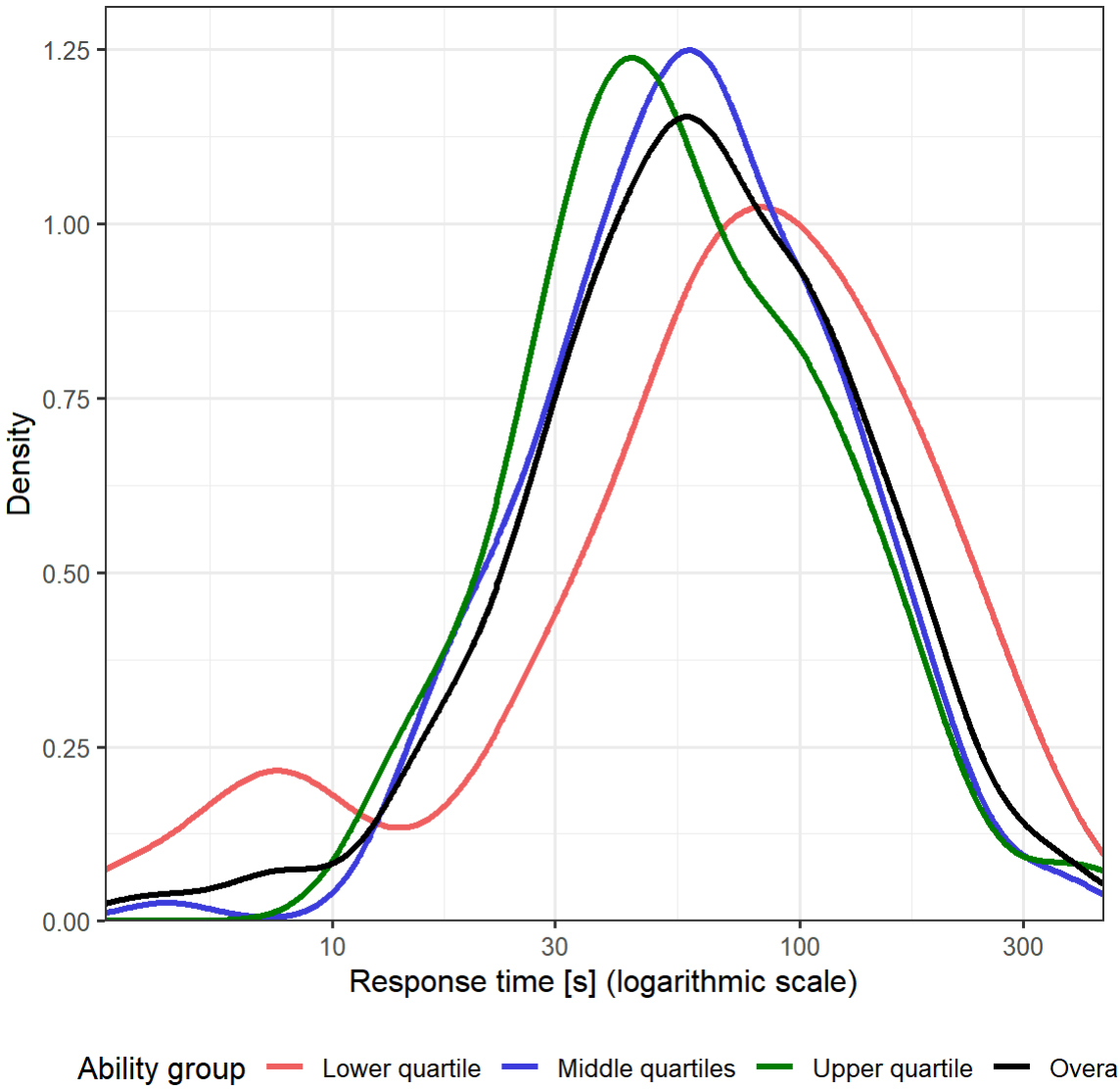
10.1.60.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.04	0.978	1.000	0.000	N/A

10.1.60.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

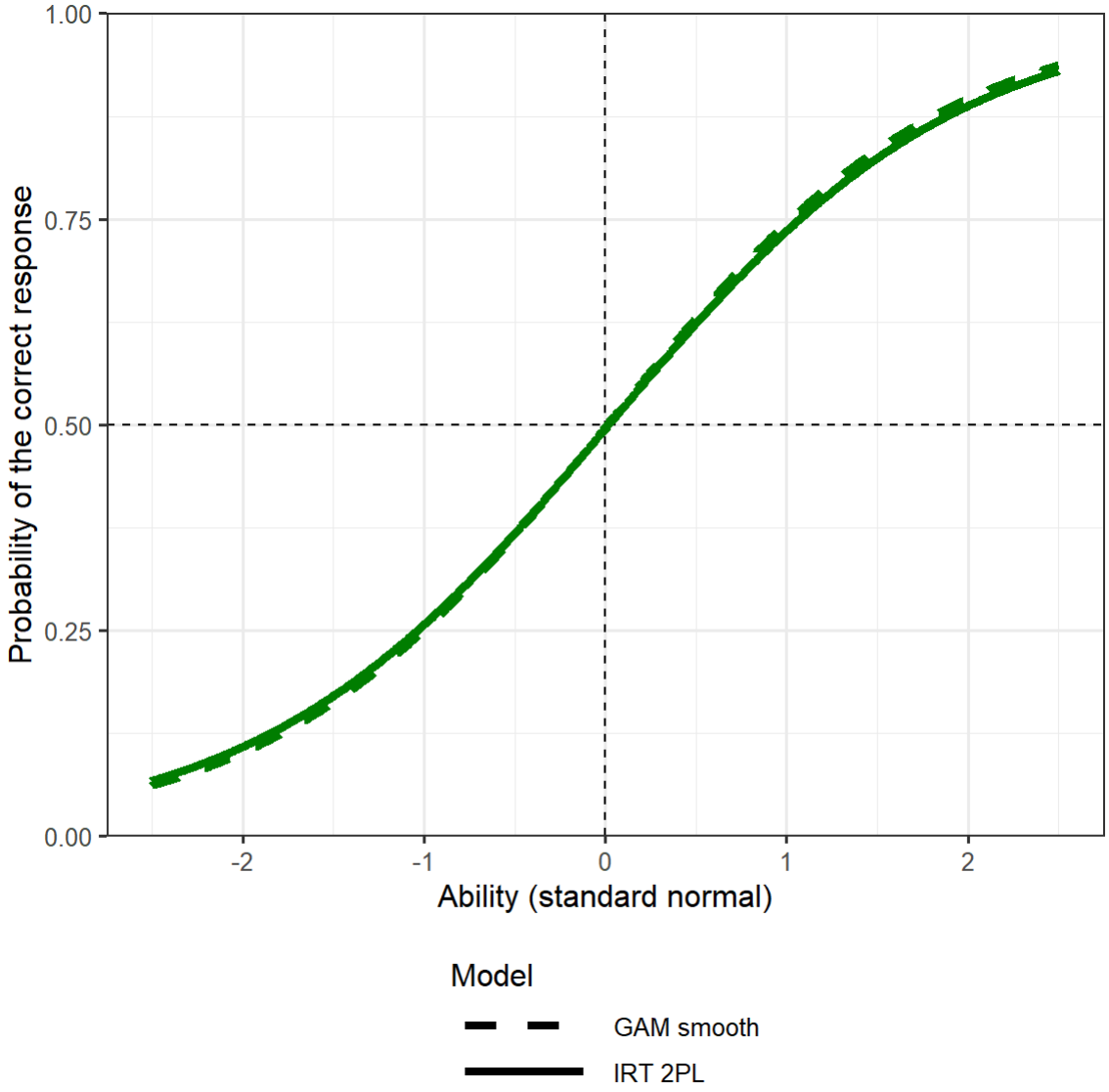
10.1.60.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.3	39.9	73.6	132.3	351.4	4.2	1.0
Middle quartiles	4.4	35.5	60.7	100.0	363.2	4.1	0.7
Upper quartile	13.2	35.4	50.7	93.6	444.4	4.0	0.7
Overall	3.3	35.6	60.5	103.9	444.4	4.1	0.8

10.1.61 M090013 (LLM)

10.1.61.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS generated items: Discrimination	IRT model including both TIMSS generated items: Item difficulty	IRT model including both TIMSS generated items: RMSE
1	49.29%	281	LLM teachers' selection	CR	0.44	0.24	1.04	0.02	0.0

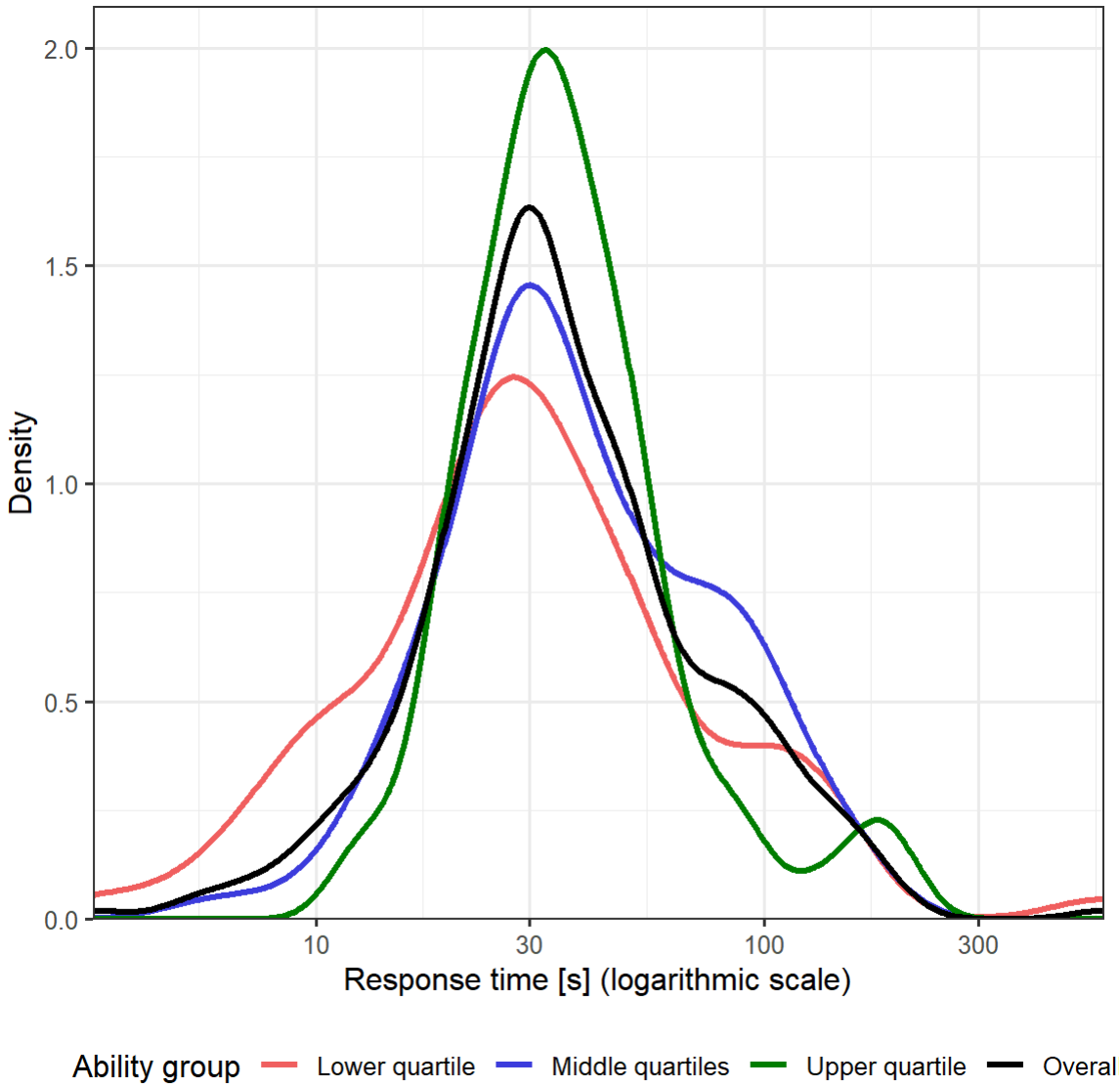
10.1.61.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.53	0.466	1.000	0.004	N/A

10.1.61.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

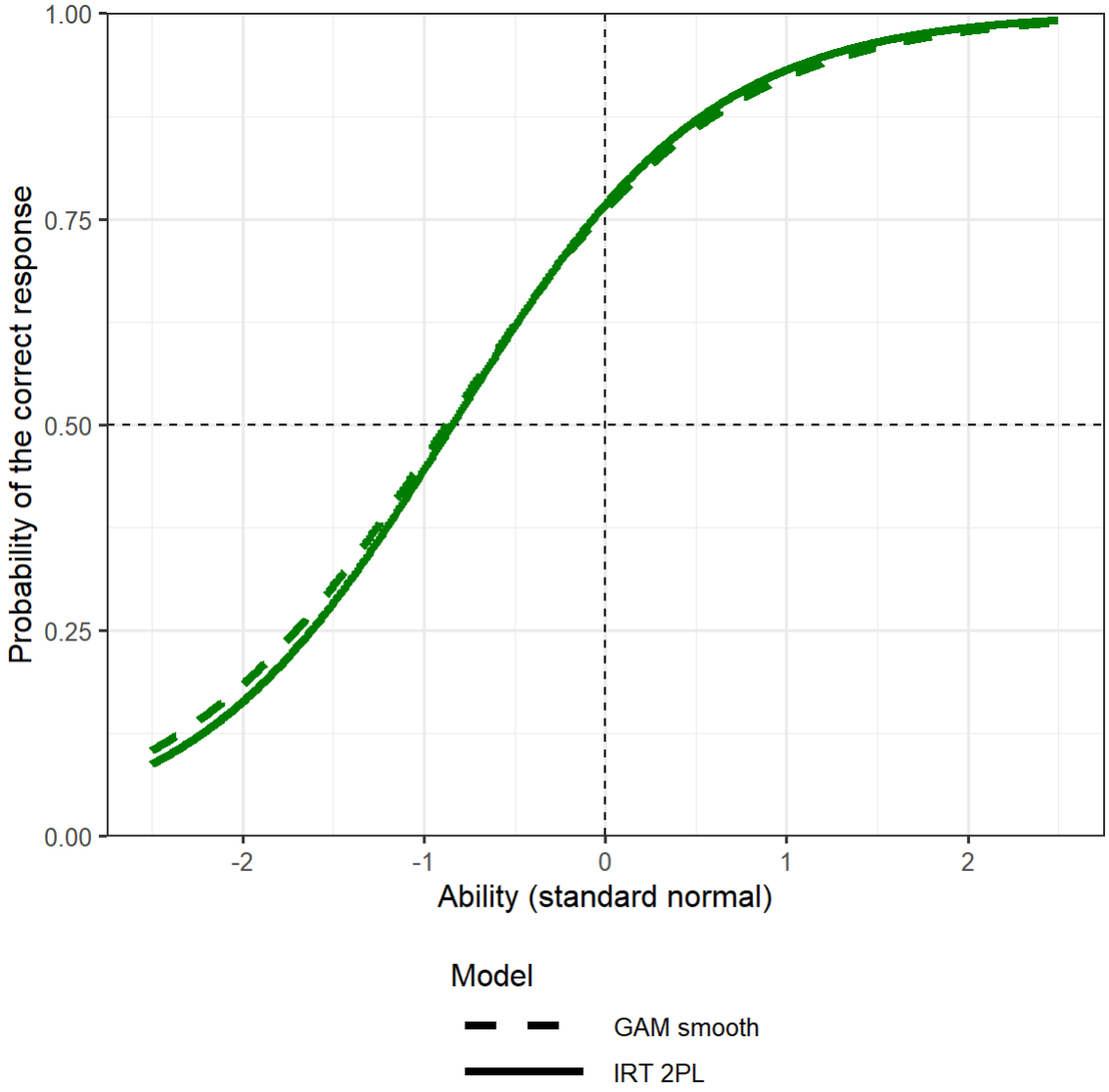
10.1.61.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.2	18.8	29.1	48.3	570.2	3.4	0.9
Middle quartiles	5.5	24.1	33.5	63.3	164.0	3.6	0.7
Upper quartile	12.3	26.7	34.5	47.5	195.0	3.6	0.6
Overall	3.2	23.5	32.3	51.3	570.2	3.6	0.7

10.1.62 M051039 (TIMSS)

10.1.62.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	69.51%	306	TIMSS	CR	0.48	0.48	1.41	−0.84	0.008

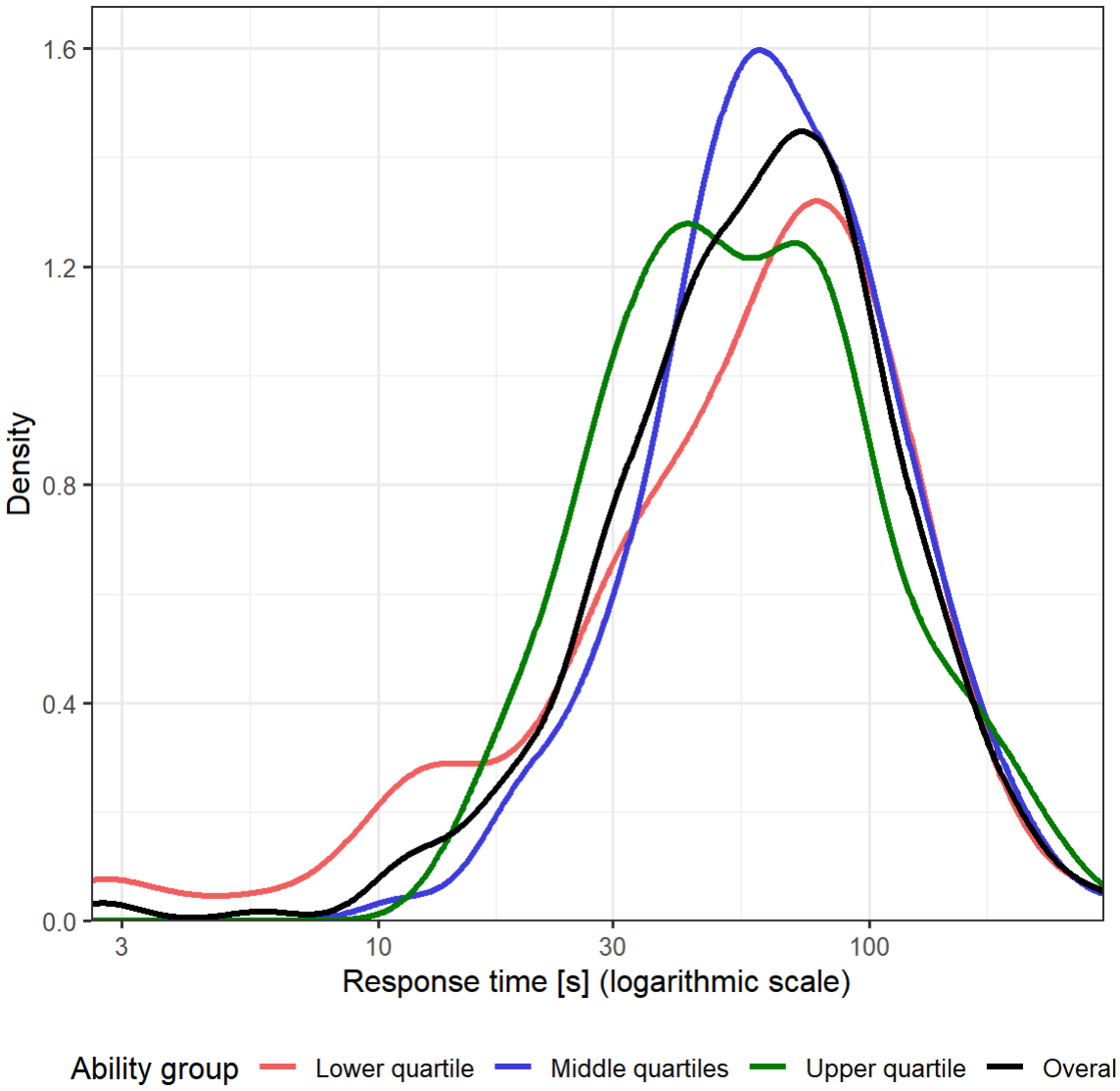
10.1.62.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.64	0.162	1.000	0.010	N/A

10.1.62.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

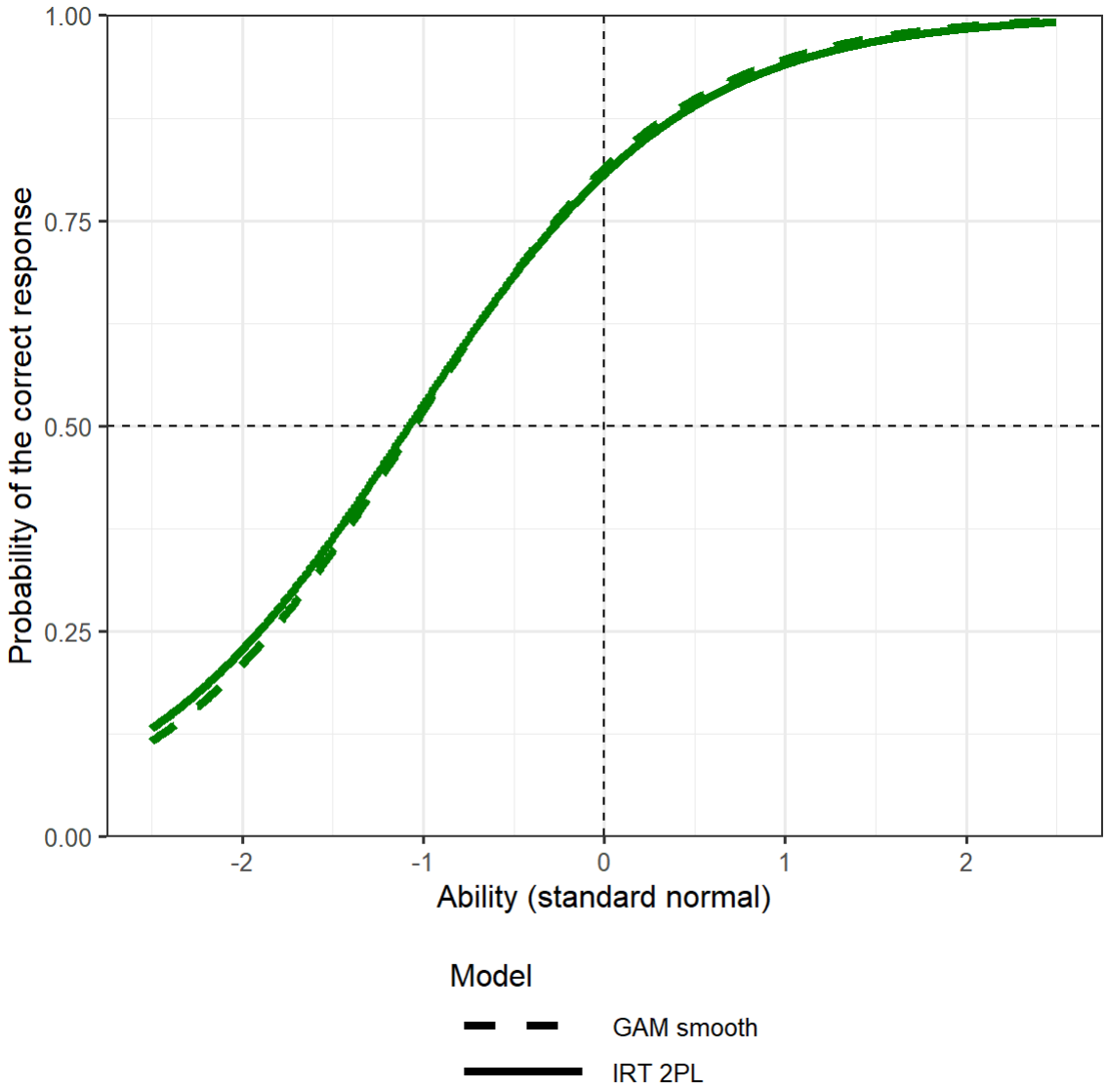
10.1.62.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.6	32.7	62.2	92.5	293.0	3.9	0.9
Middle quartiles	11.1	44.5	63.4	92.6	298.8	4.1	0.6
Upper quartile	16.6	34.6	51.9	83.3	258.3	4.0	0.6
Overall	2.6	39.3	61.6	89.5	298.8	4.0	0.7

10.1.63 M100011 (LLM)

10.1.63.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	74.16%	299	LLM simple setup	CR	0.47	0.25	1.33	−1.08	0.016

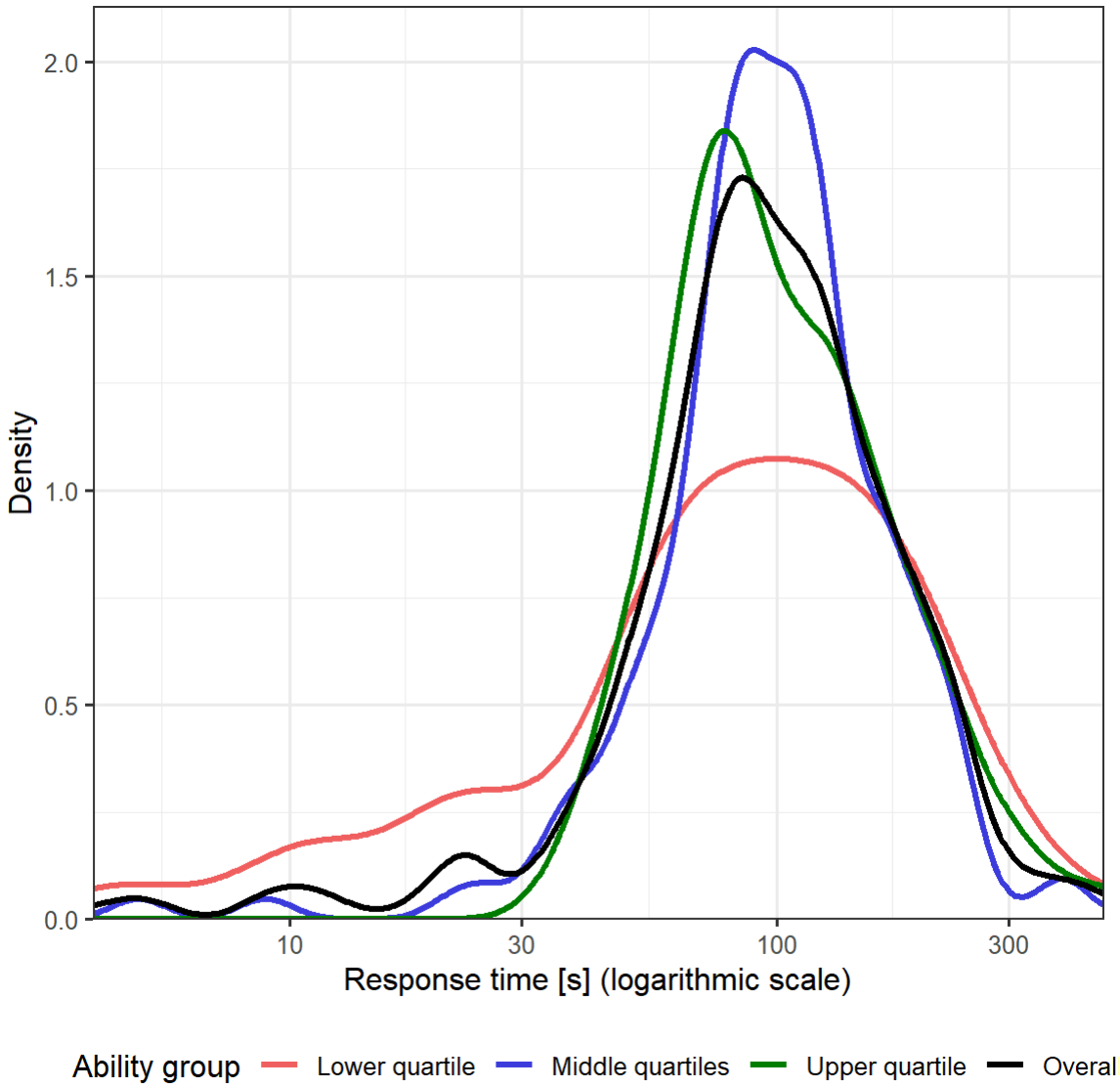
10.1.63.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.65	0.724	1.000	0.002	N/A

10.1.63.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

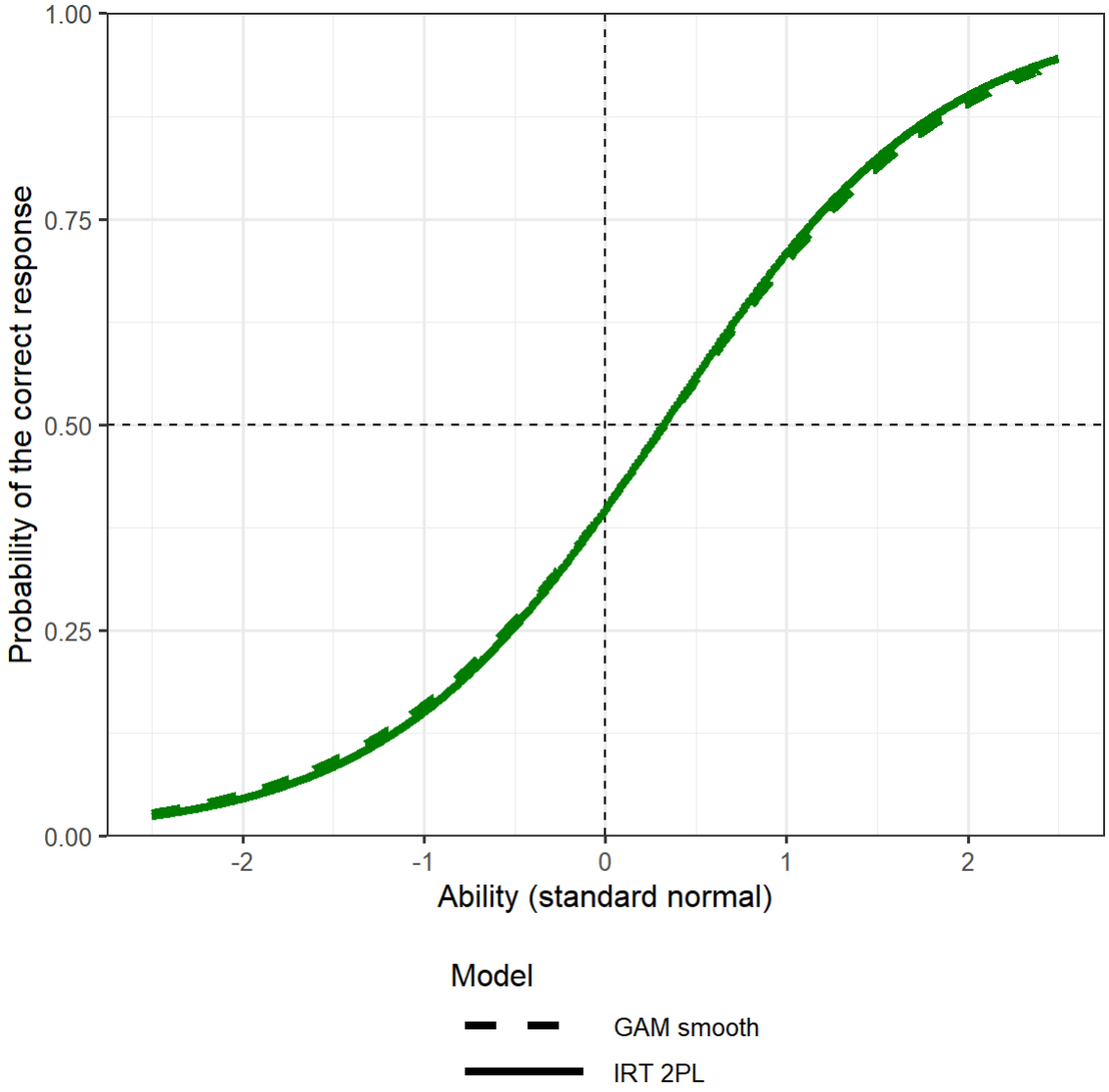
10.1.63.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.0	53.7	88.0	149.5	437.9	4.3	1.0
Middle quartiles	4.9	75.3	98.6	128.0	392.3	4.6	0.6
Upper quartile	37.8	70.7	92.1	135.9	468.2	4.6	0.5
Overall	4.0	69.6	94.9	134.7	468.2	4.5	0.7

10.1.64 M890006 (LLM)

10.1.64.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS generated items: Discrimination	IRT model including both TIMSS generated items: Item difficulty	IRT model including both TIMSS generated items: RMSE
1	41.79%	281	LLM complex setup	CR	0.47	0.30	1.31	0.32	0.00

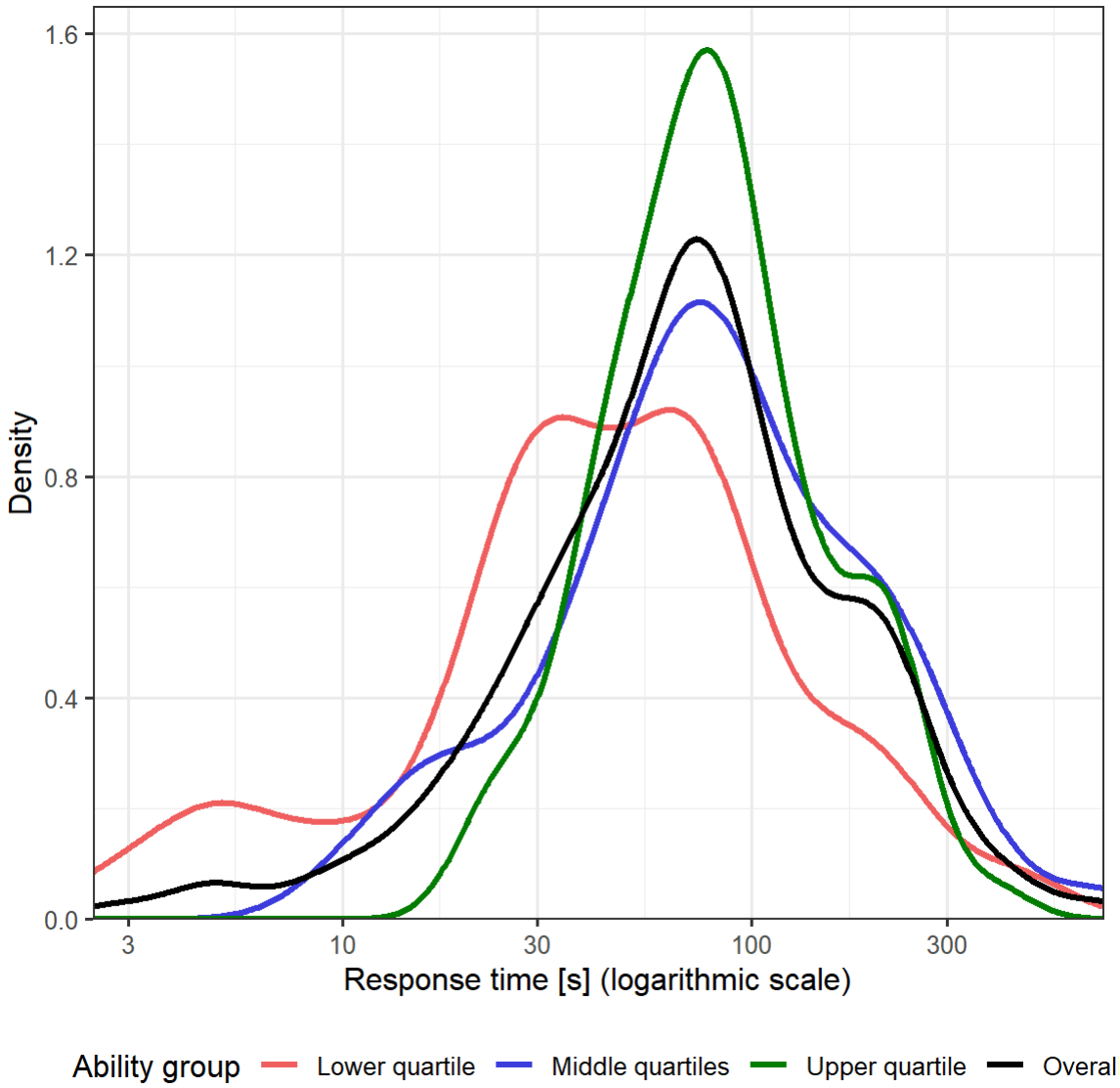
10.1.64.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.02	0.364	1.000	0.006	N/A

10.1.64.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

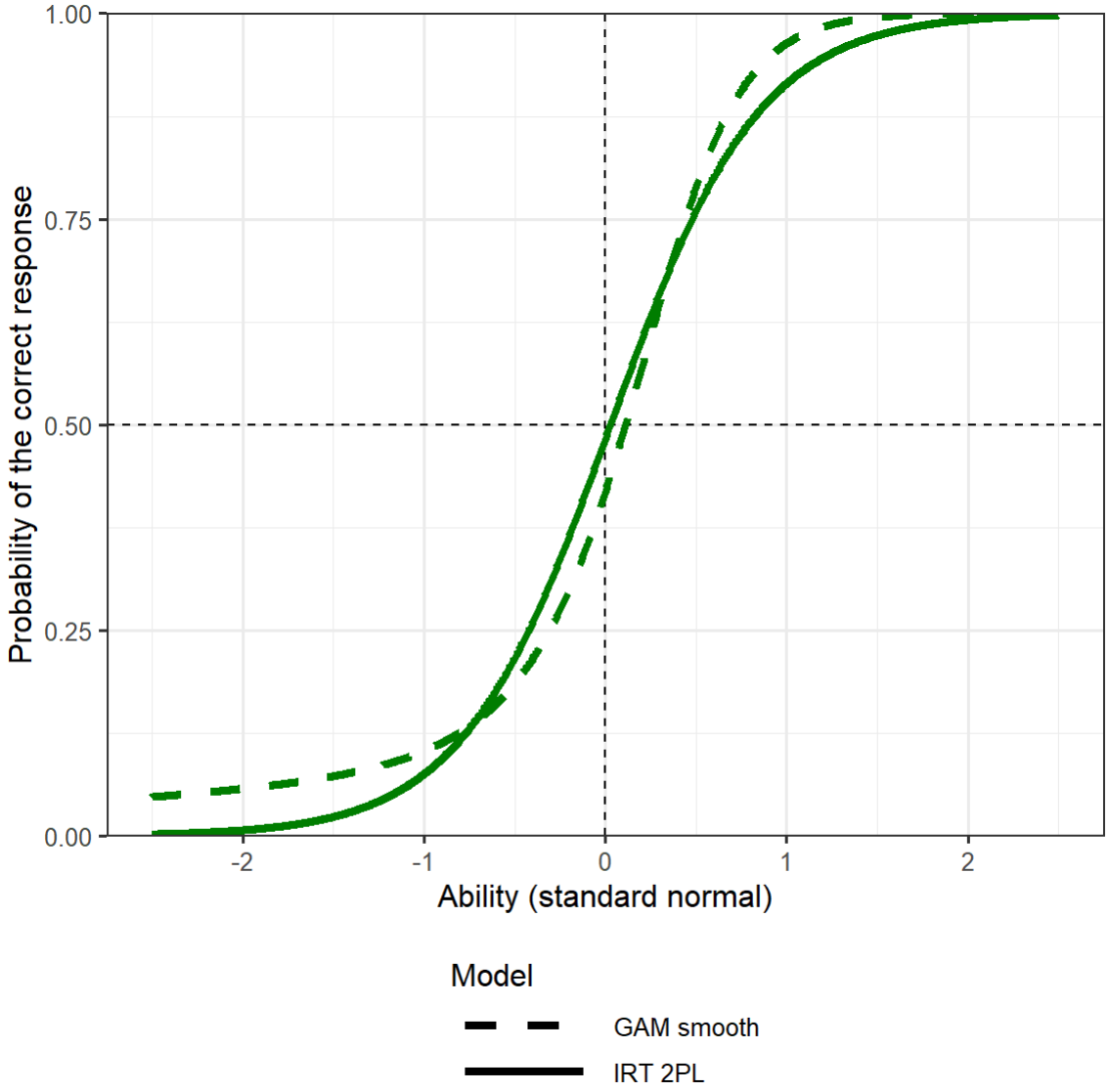
10.1.64.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.5	25.4	44.0	78.6	444.2	3.7	1.1
Middle quartiles	9.1	46.1	79.1	144.9	723.8	4.3	0.9
Upper quartile	21.4	54.0	81.1	114.0	387.9	4.4	0.6
Overall	2.5	39.7	70.9	114.3	723.8	4.2	0.9

10.1.65 M061039 (TIMSS)

10.1.65.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	50.67%	299	TIMSS	CR	0.67	0.72	2.44	0.03	0.023

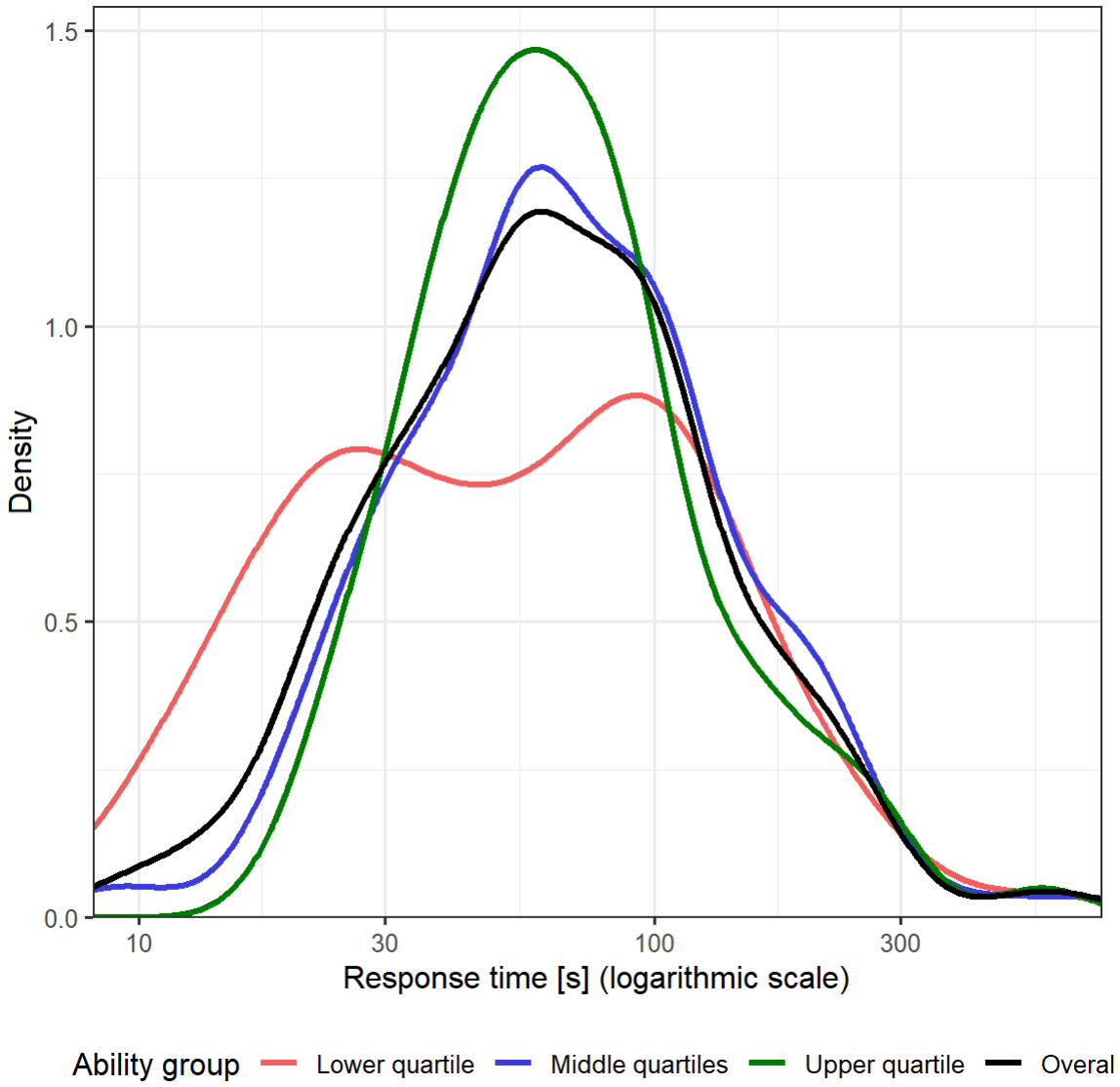
10.1.65.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.63	0.268	1.000	0.006	N/A

10.1.65.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

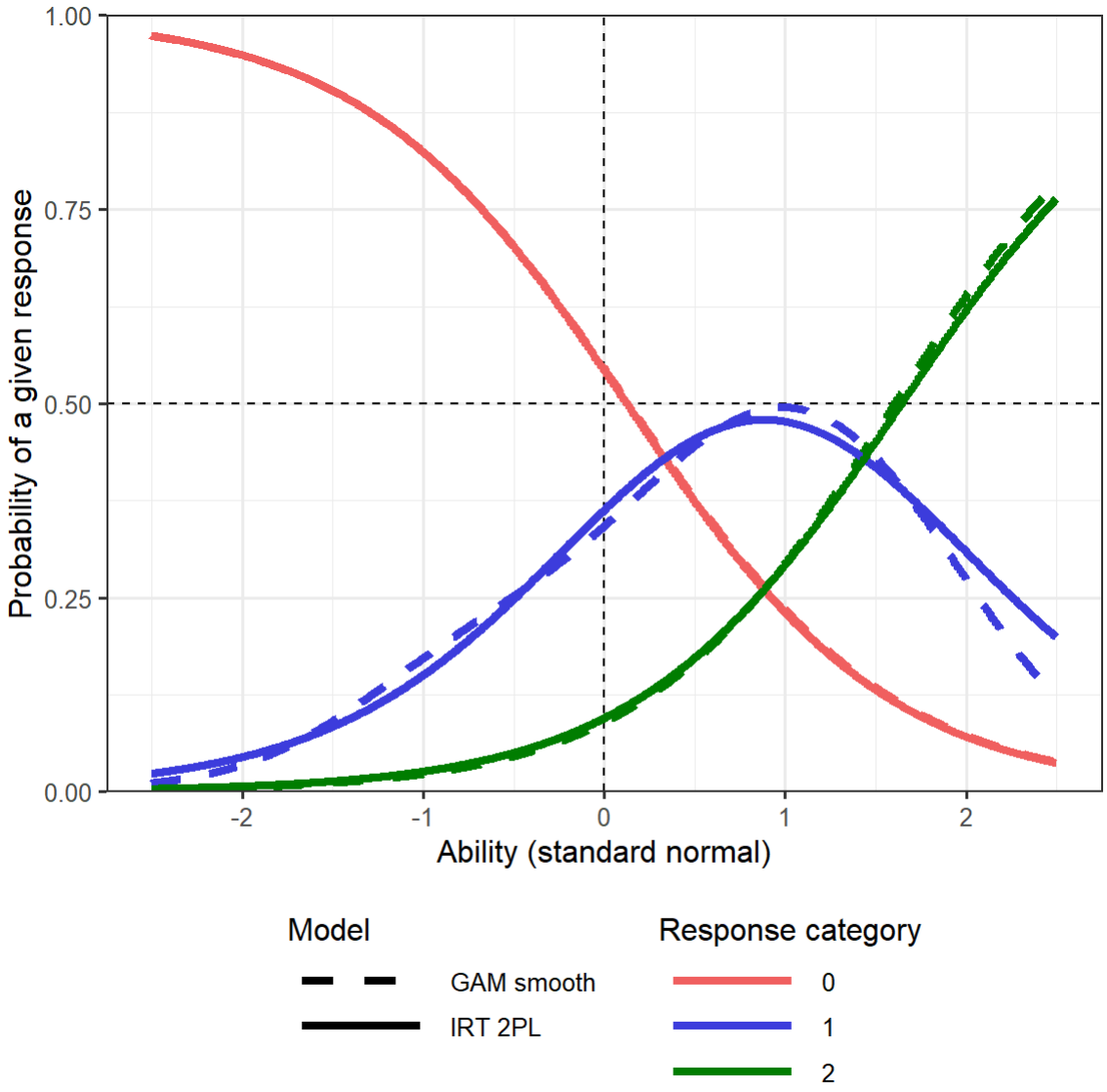
10.1.65.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	8.7	24.7	50.6	103.8	591.2	3.9	0.9
Middle quartiles	8.1	41.6	62.4	104.7	734.9	4.2	0.7
Upper quartile	19.7	41.4	61.1	91.0	568.4	4.2	0.6
Overall	8.1	36.5	61.3	100.2	734.9	4.1	0.8

10.1.66 M100002 (LLM)

10.1.66.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMS Discrimination	IRT model including both TIMS Item difficulty	IRT model including both TIMS Catego difficul
2	31.35%	319	LLM simple setup	CR	0.56	0.41	1.37	0.88	0.1

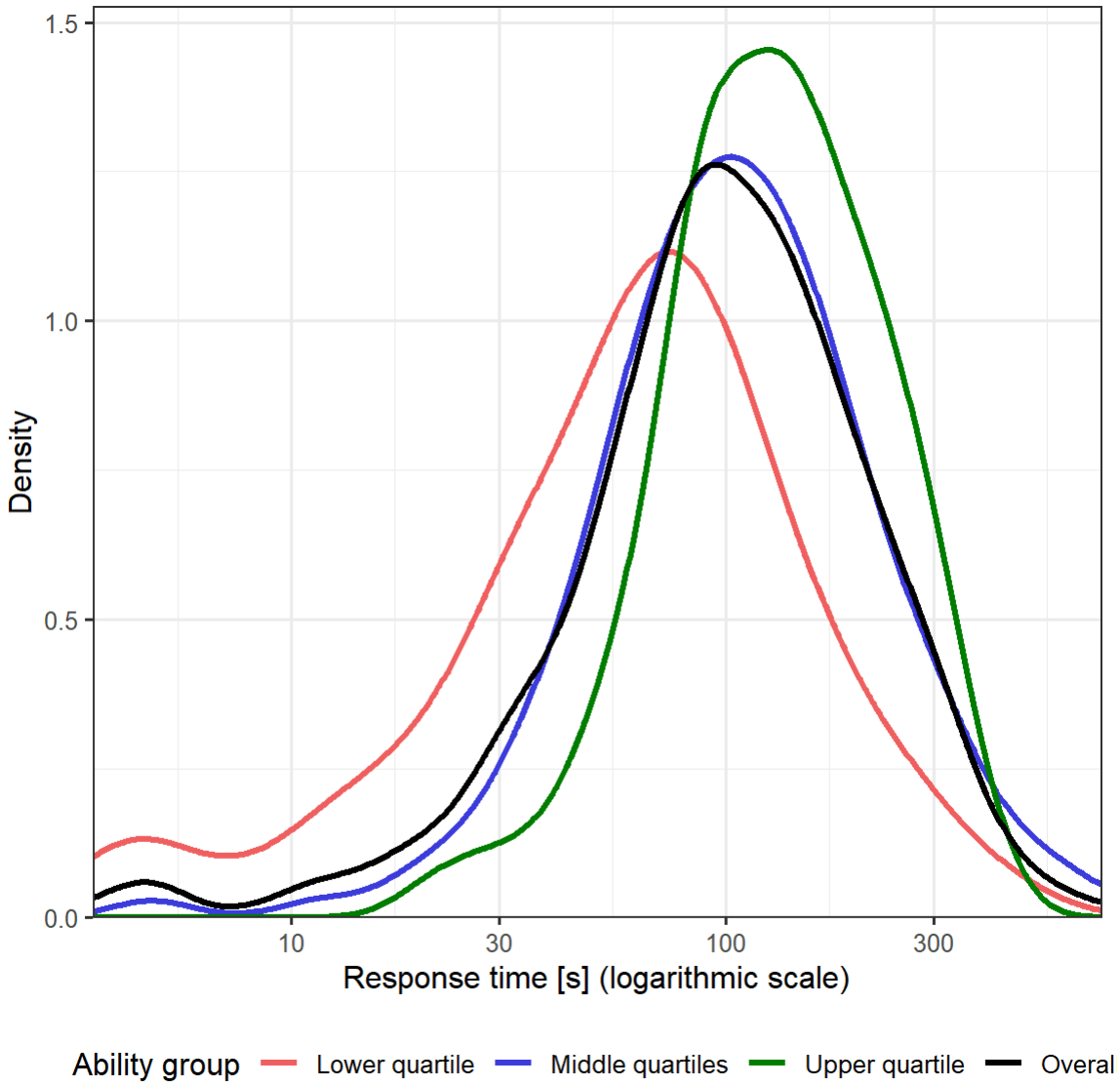
10.1.66.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.02	0.990	1.000	0.000	N/A

10.1.66.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

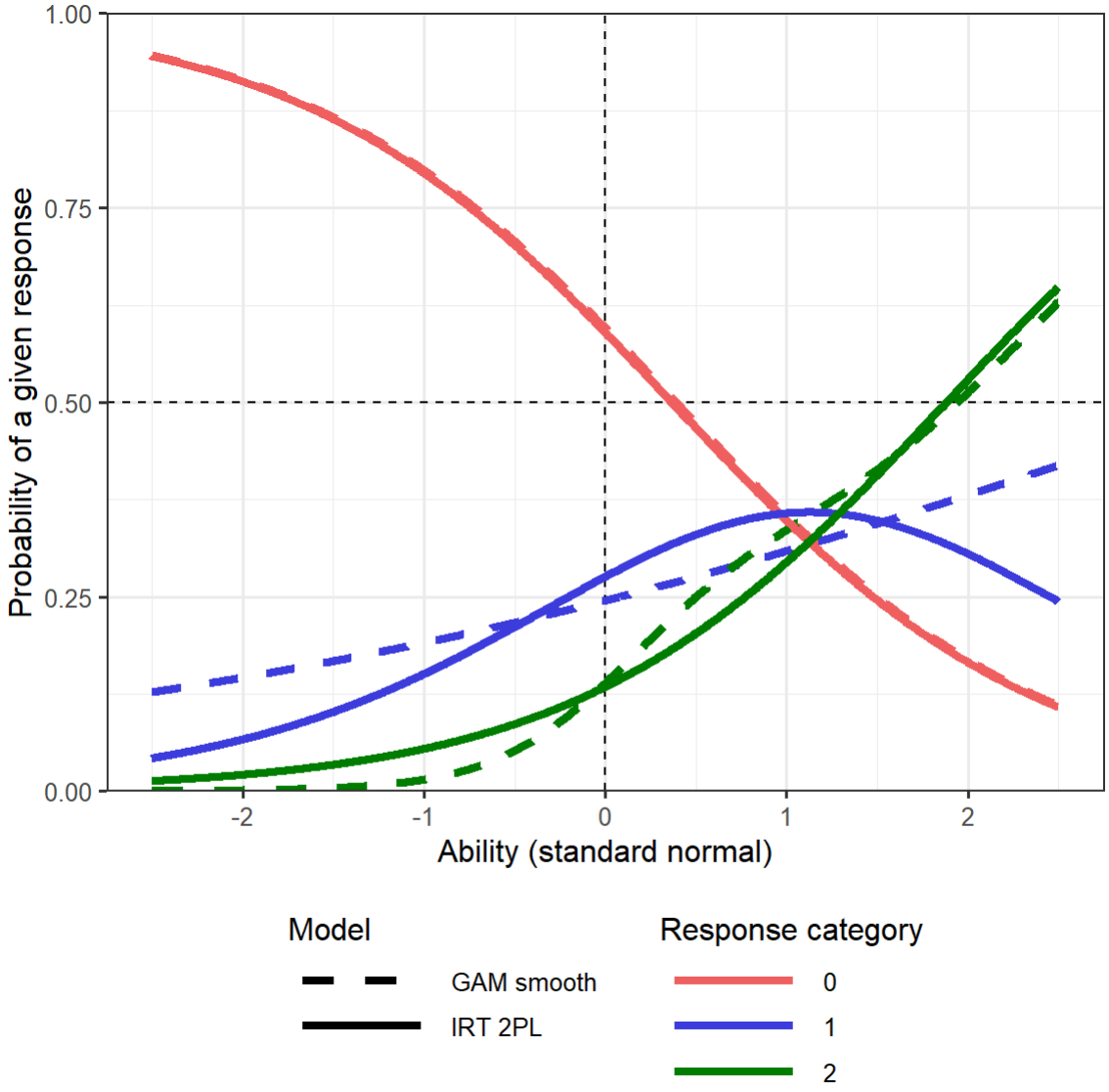
10.1.66.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	34.6	68.8	107.8	443.4	4.1	1.0
Middle quartiles	4.8	66.9	102.1	163.7	730.3	4.6	0.8
Upper quartile	21.5	88.3	133.1	196.1	389.0	4.8	0.6
Overall	3.5	61.9	98.7	159.8	730.3	4.5	0.8

10.1.67 M090047 (LLM)

10.1.67.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIM Discrimination	IRT model including both TIM Item difficulty	IRT model including both TIM Categ difficulty
2	29.19%	300	LLM teachers' selection	CR	0.46	0.29	0.99	1.13	(

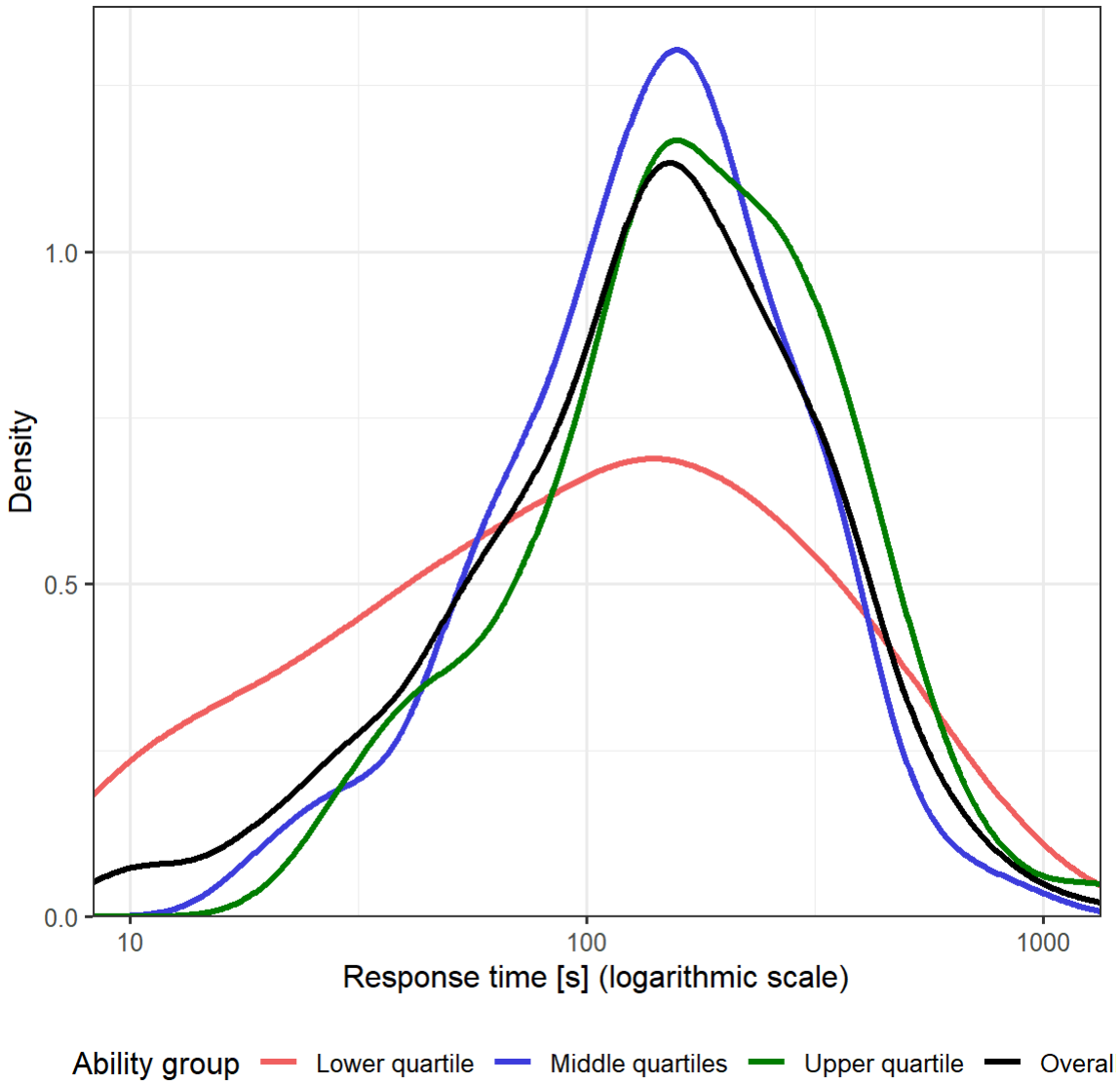
10.1.67.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.87	0.238	1.000	0.008	N/A

10.1.67.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

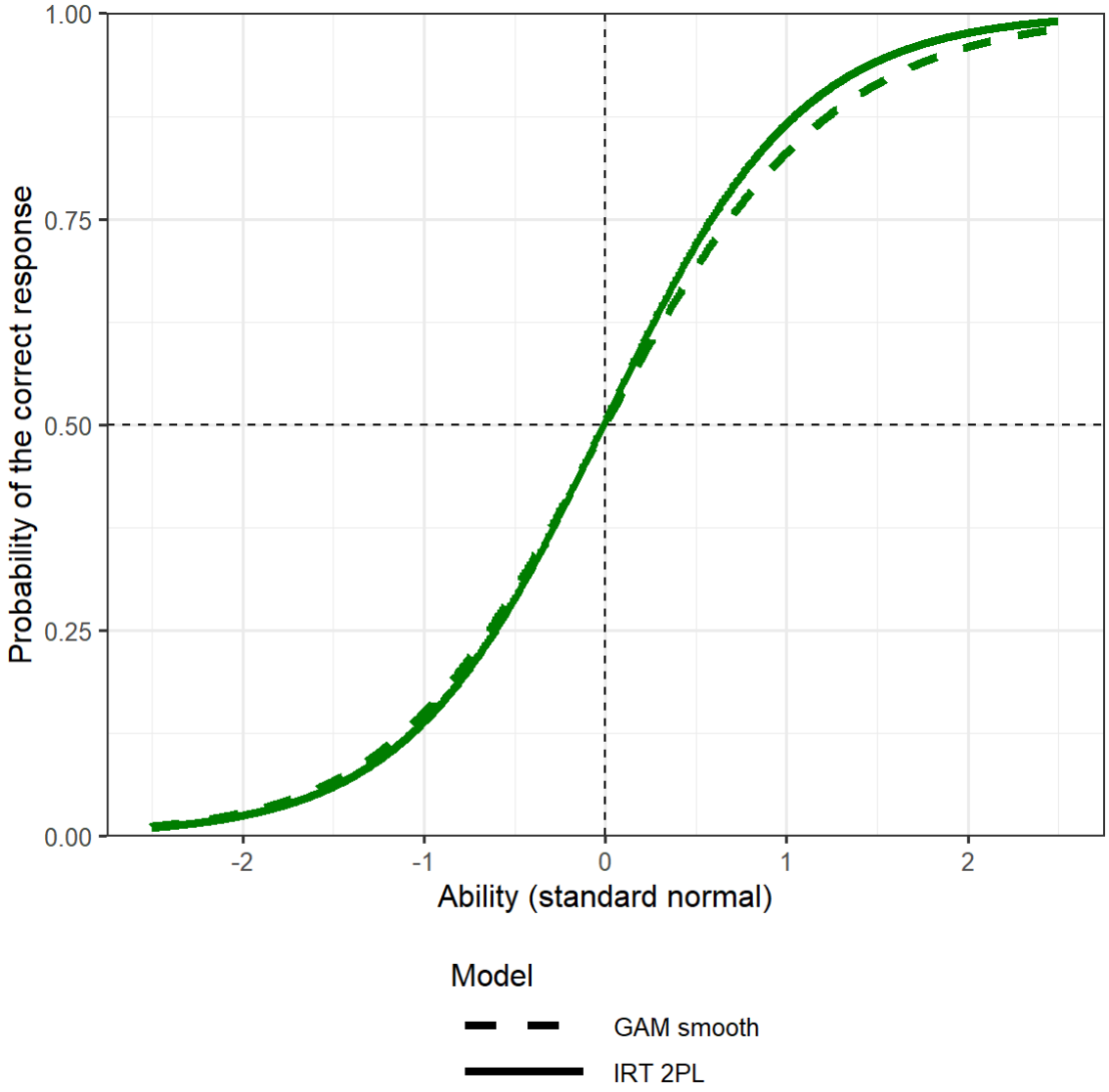
10.1.67.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	8.3	38.2	114.5	224.7	914.0	4.5	1.2
Middle quartiles	18.6	83.2	148.1	217.7	922.5	4.9	0.7
Upper quartile	25.8	104.1	163.6	268.6	1,335.1	5.1	0.8
Overall	8.3	78.2	145.9	241.8	1,335.1	4.9	0.9

10.1.68 M100013 (LLM)

10.1.68.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	51.32%	304	LLM simple setup	CR	0.55	0.36	1.84	−0.01	0.018

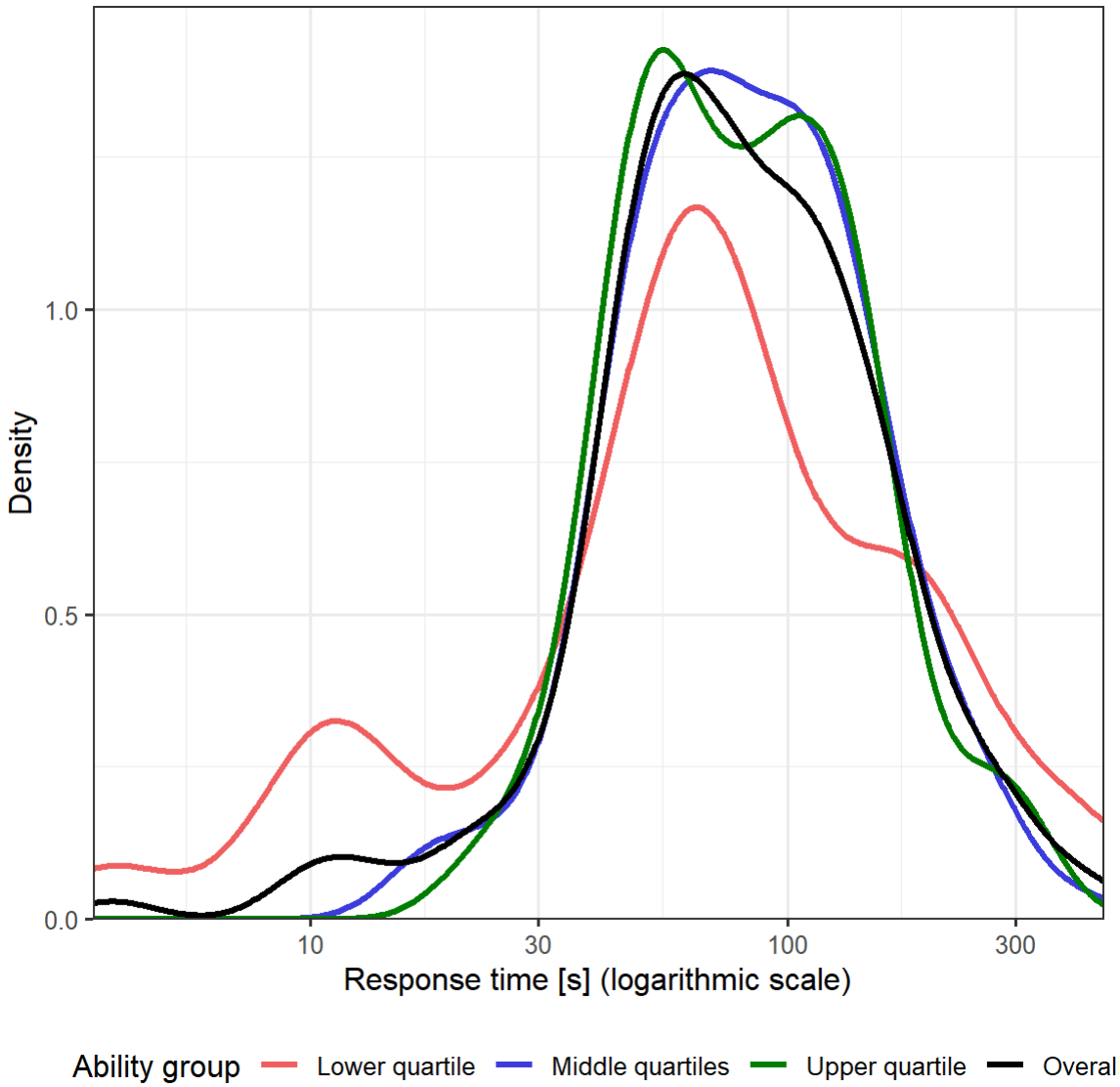
10.1.68.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.55	0.758	1.000	0.001	N/A

10.1.68.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

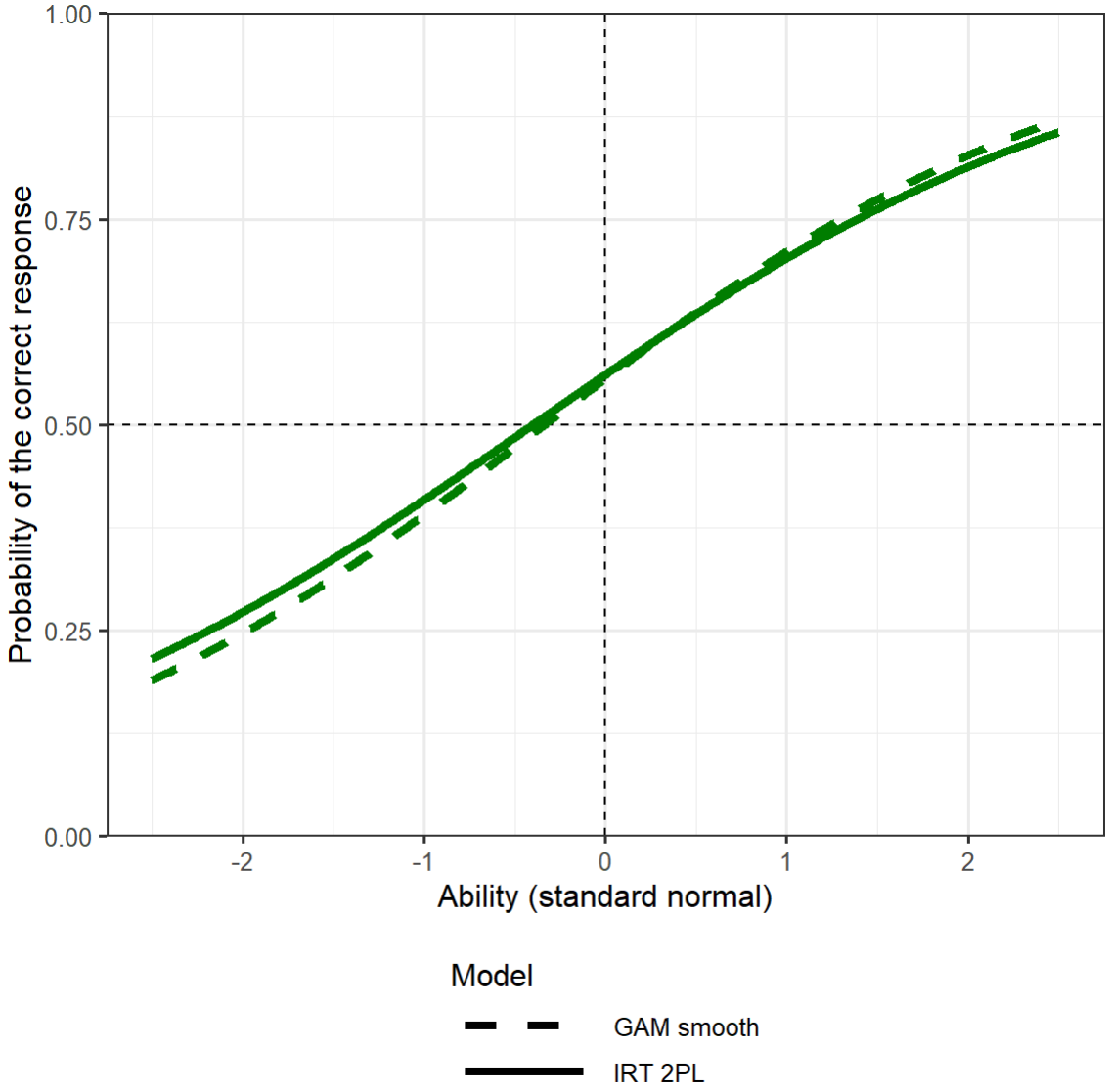
10.1.68.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	43.3	64.0	120.5	458.5	4.1	1.1
Middle quartiles	15.7	53.8	80.3	122.2	416.2	4.4	0.6
Upper quartile	22.0	51.8	77.0	120.8	323.9	4.4	0.6
Overall	3.5	51.7	75.7	121.2	458.5	4.3	0.7

10.1.69 M100007 (LLM)

10.1.69.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS a generated items: Discrimination	Item difficulty	RMSEA
1	55.92%	304	LLM simple setup	CR	0.30	0.18	0.61	−0.40	0.000

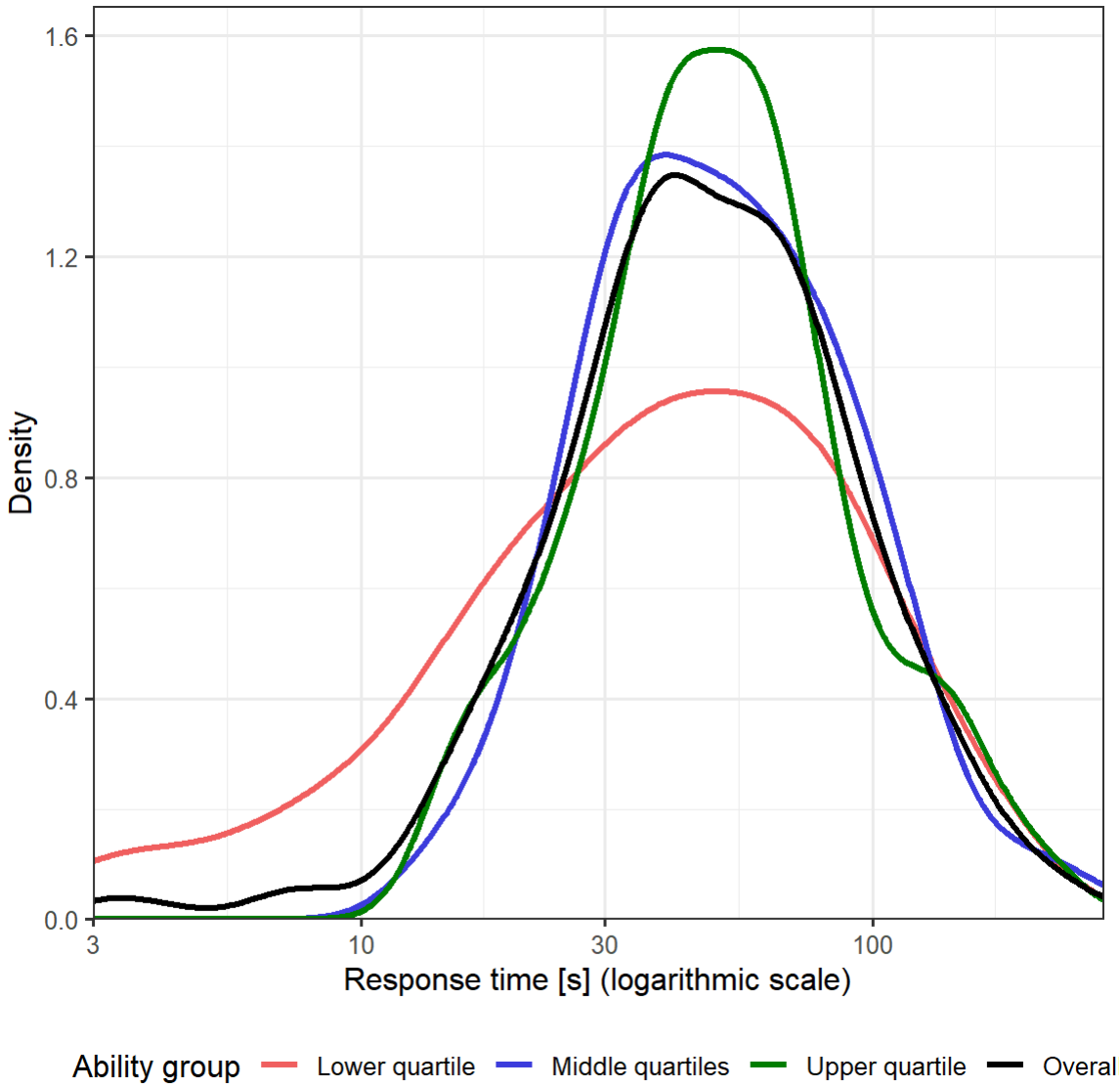
10.1.69.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.04	0.979	1.000	0.000	N/A

10.1.69.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

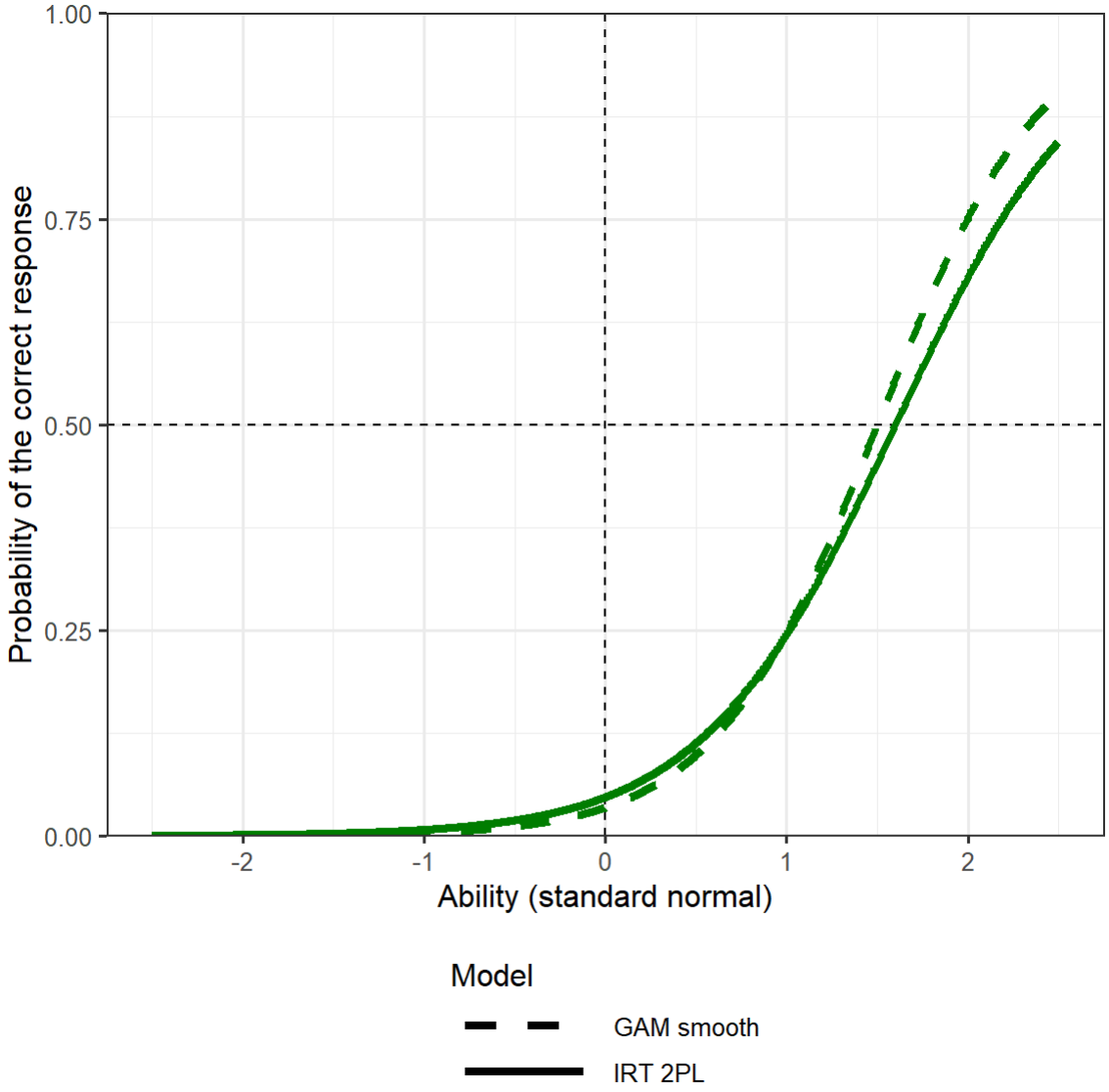
10.1.69.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	21.0	39.1	69.9	165.7	3.6	0.9
Middle quartiles	13.1	32.9	50.2	74.2	281.3	3.9	0.6
Upper quartile	14.9	35.3	48.3	68.6	225.3	3.9	0.6
Overall	3.0	31.2	46.1	73.8	281.3	3.8	0.7

10.1.70 M090042 (LLM)

10.1.70.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	12.12%	298	LLM teachers' selection	CR	0.47	0.26	1.88	1.60	0.0

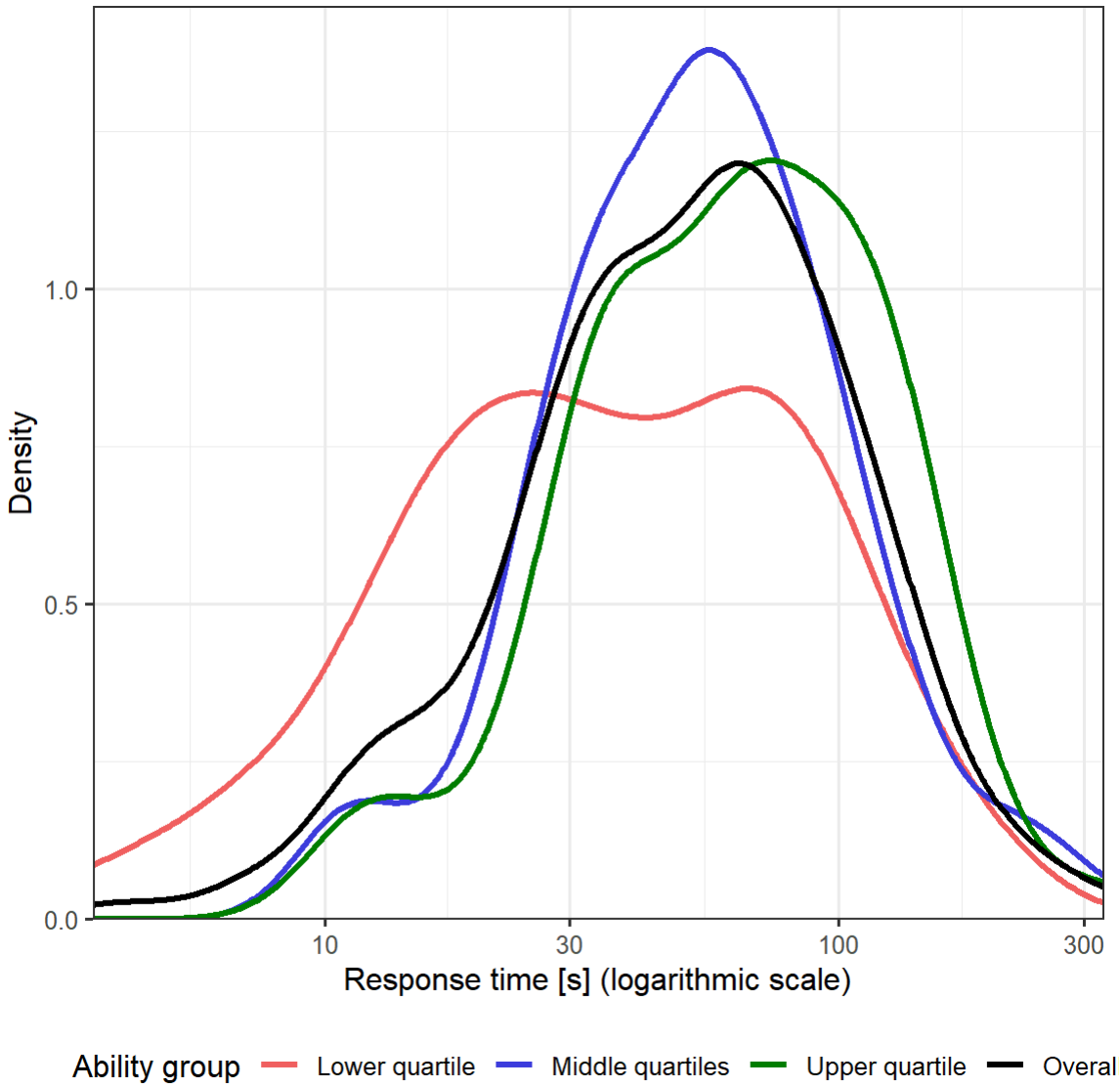
10.1.70.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
5.37	0.068	1.000	0.025	N/A

10.1.70.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

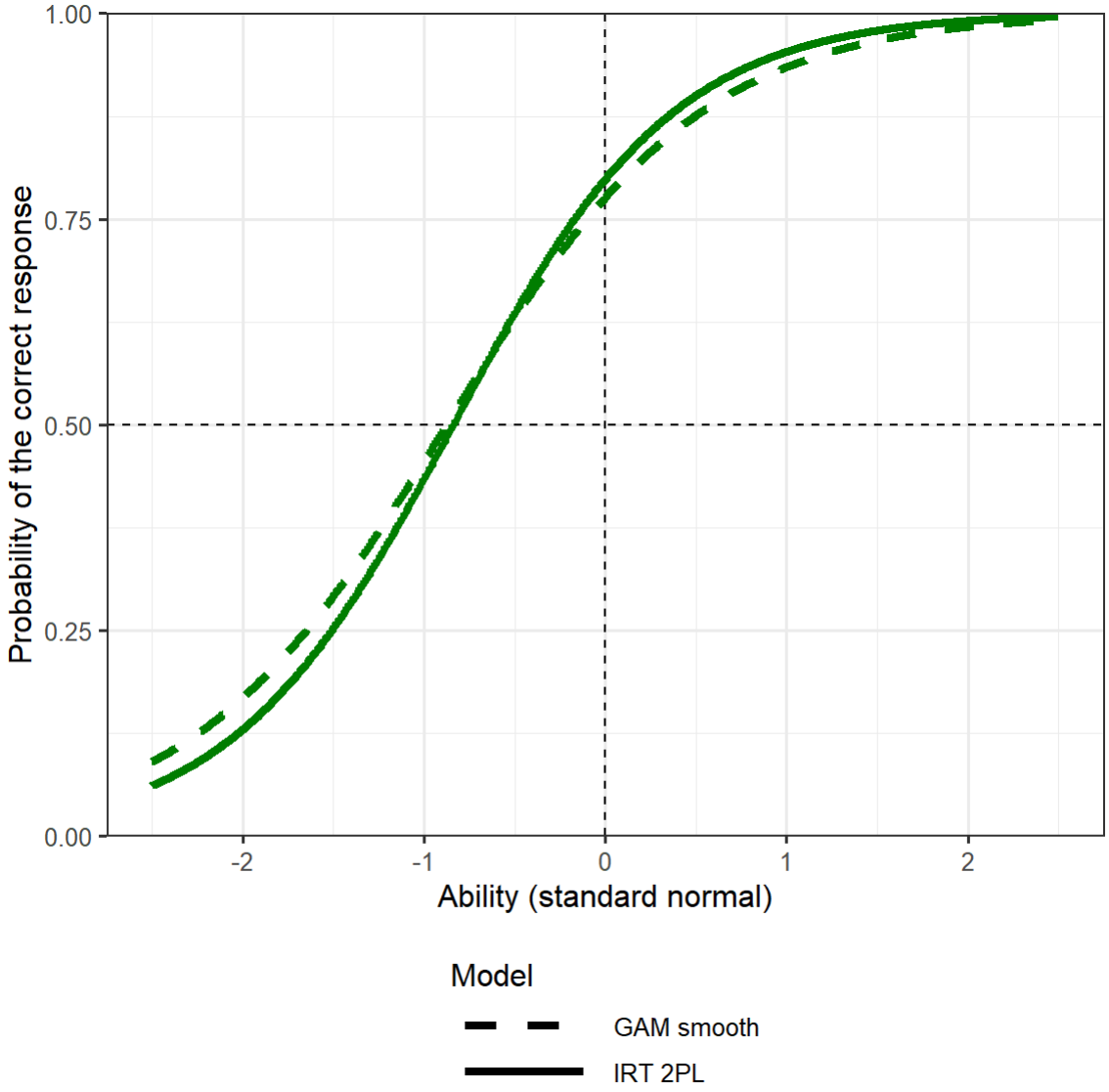
10.1.70.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	18.0	32.4	70.8	199.2	3.5	0.9
Middle quartiles	9.8	35.1	53.3	83.0	323.6	4.0	0.7
Upper quartile	11.0	37.1	64.0	103.2	326.9	4.1	0.7
Overall	3.5	30.8	52.5	84.8	326.9	3.9	0.8

10.1.71 M041174 (TIMSS)

10.1.71.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	71.38%	304	TIMSS	CR	0.49	0.43	1.64	−0.84	0.020

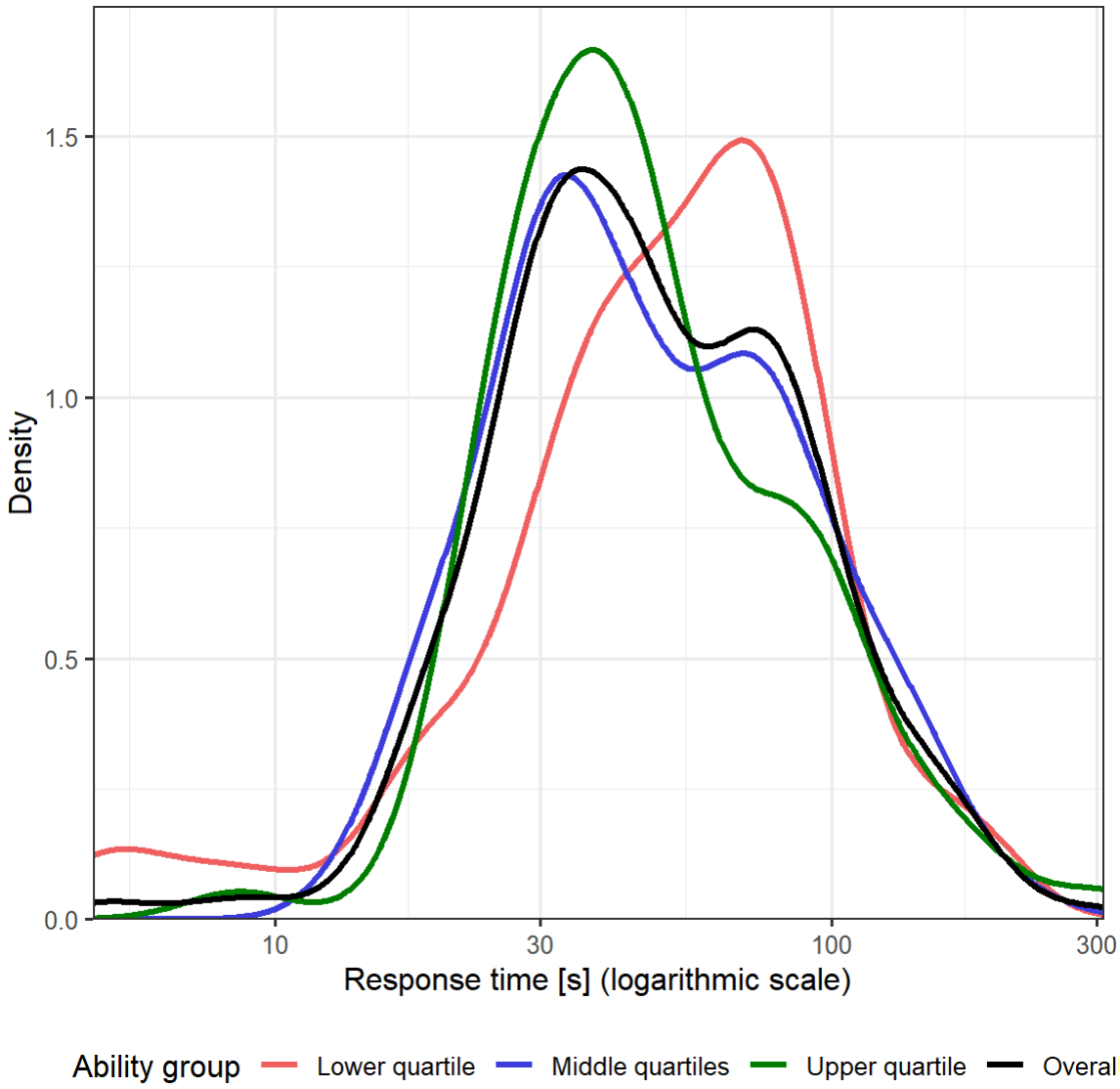
10.1.71.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.16	0.339	1.000	0.006	N/A

10.1.71.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

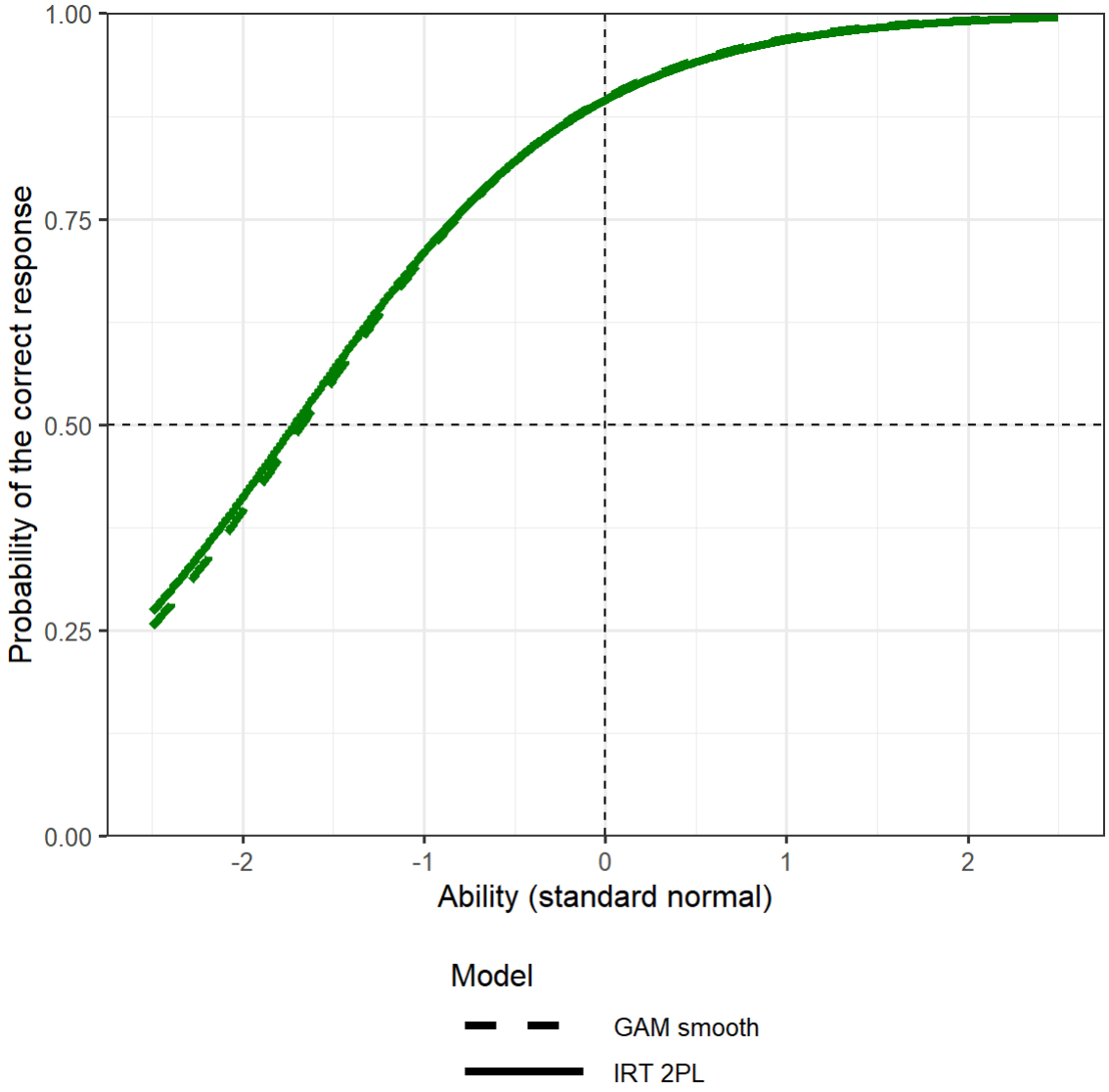
10.1.71.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.7	34.7	52.5	77.0	181.4	3.9	0.7
Middle quartiles	13.4	30.6	44.3	75.0	236.4	3.9	0.6
Upper quartile	8.7	30.9	41.5	73.5	307.8	3.8	0.6
Overall	4.7	30.8	45.3	76.1	307.8	3.9	0.6

10.1.72 M890011 (LLM)

10.1.72.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	84.33%	319	LLM complex setup	CR	0.38	0.35	1.25	−1.71	0.01

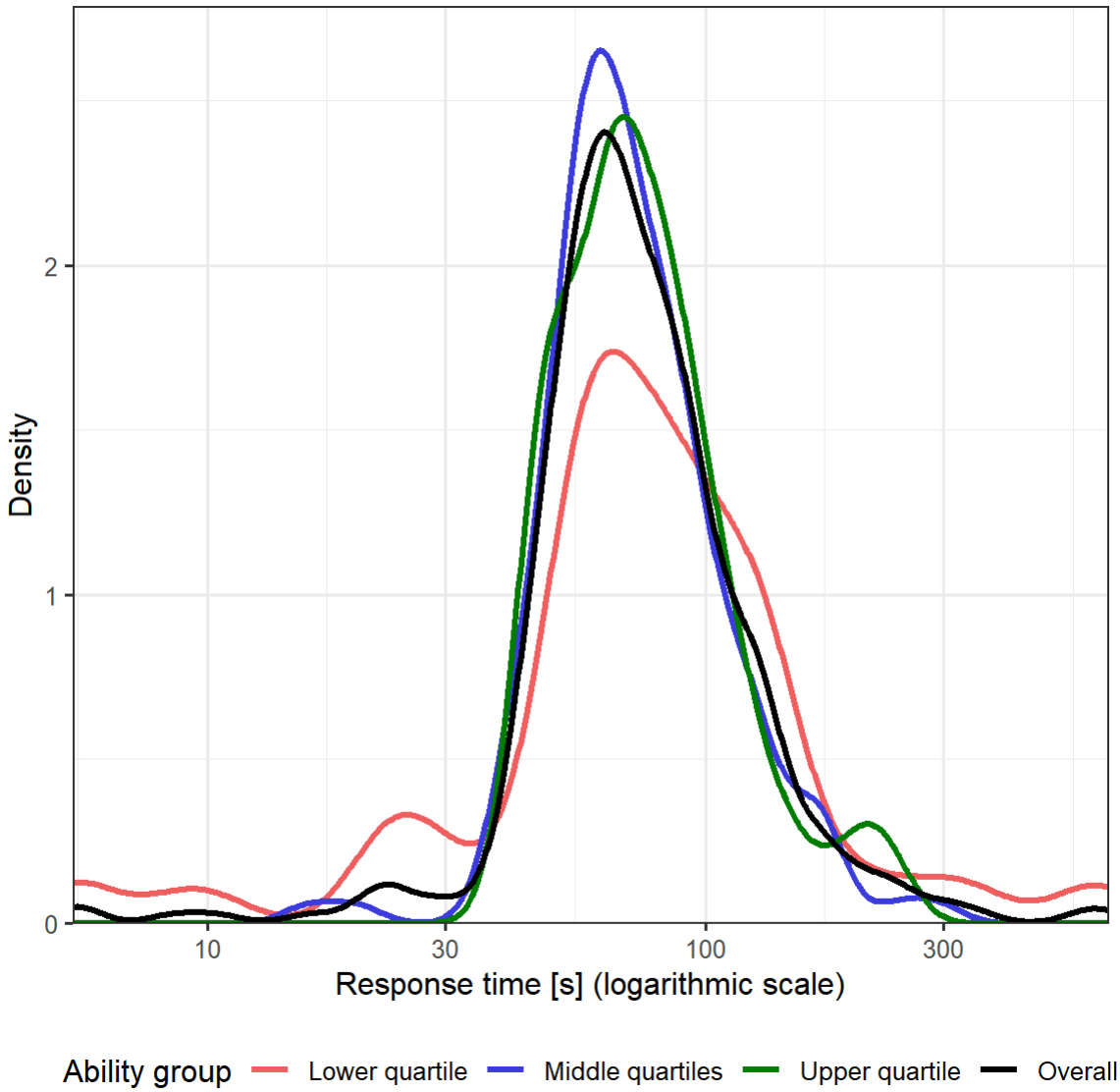
10.1.72.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.91	0.384	1.000	0.007	N/A

10.1.72.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

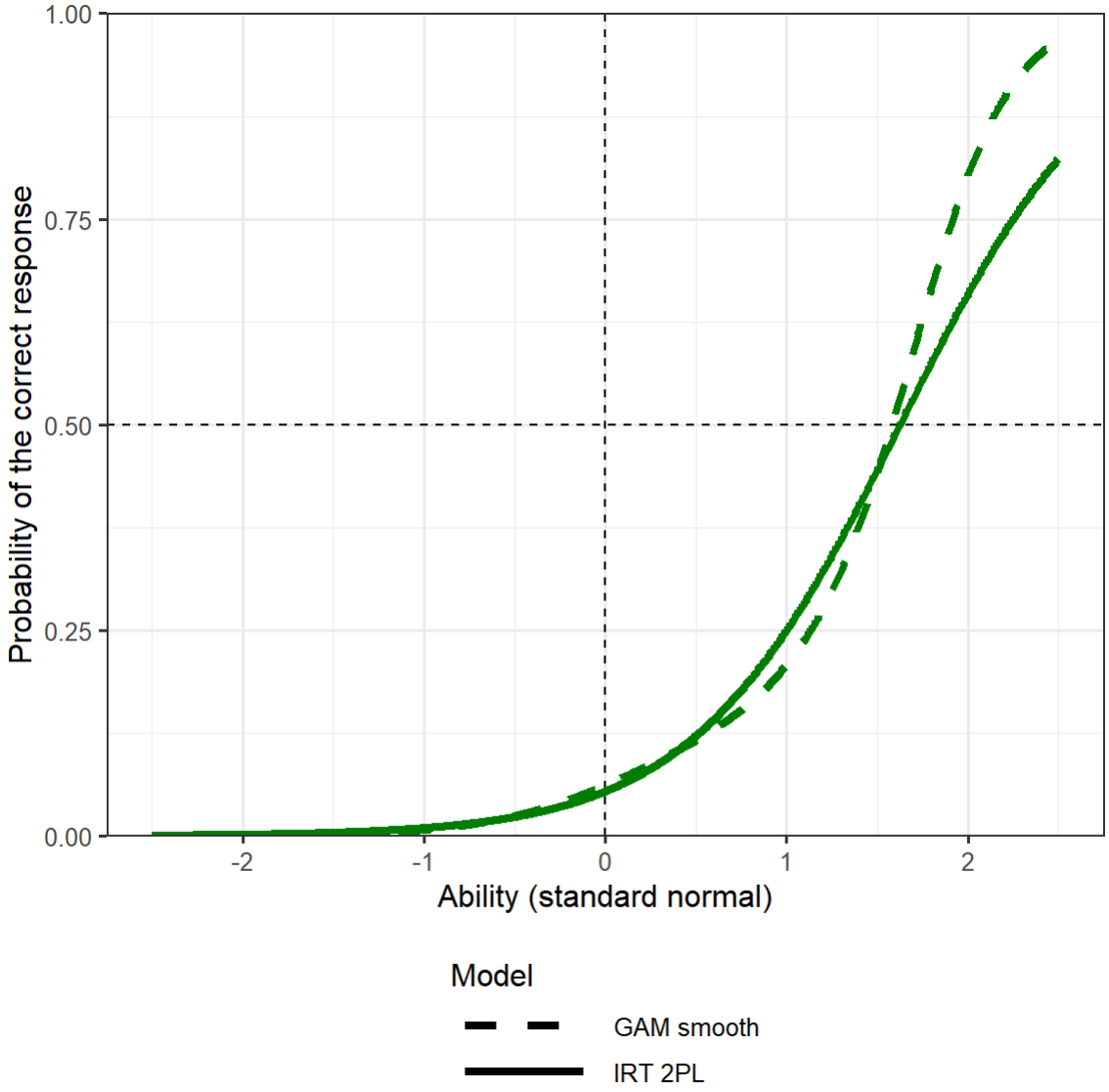
10.1.72.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.3	56.1	75.1	108.9	643.9	4.3	0.8
Middle quartiles	15.9	55.7	67.8	89.0	294.0	4.3	0.4
Upper quartile	42.0	57.4	70.4	91.0	237.8	4.3	0.4
Overall	5.3	56.1	70.4	94.7	643.9	4.3	0.5

10.1.73 M100017 (LLM)

10.1.73.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS a generated items: Discrimination	IRT model including both TIMSS a generated items: Item difficulty	IRT model including both TIMSS a generated items: RMSEA
1	12.42%	299	LLM simple setup	CR	0.45	0.25	1.76	1.63	0.021

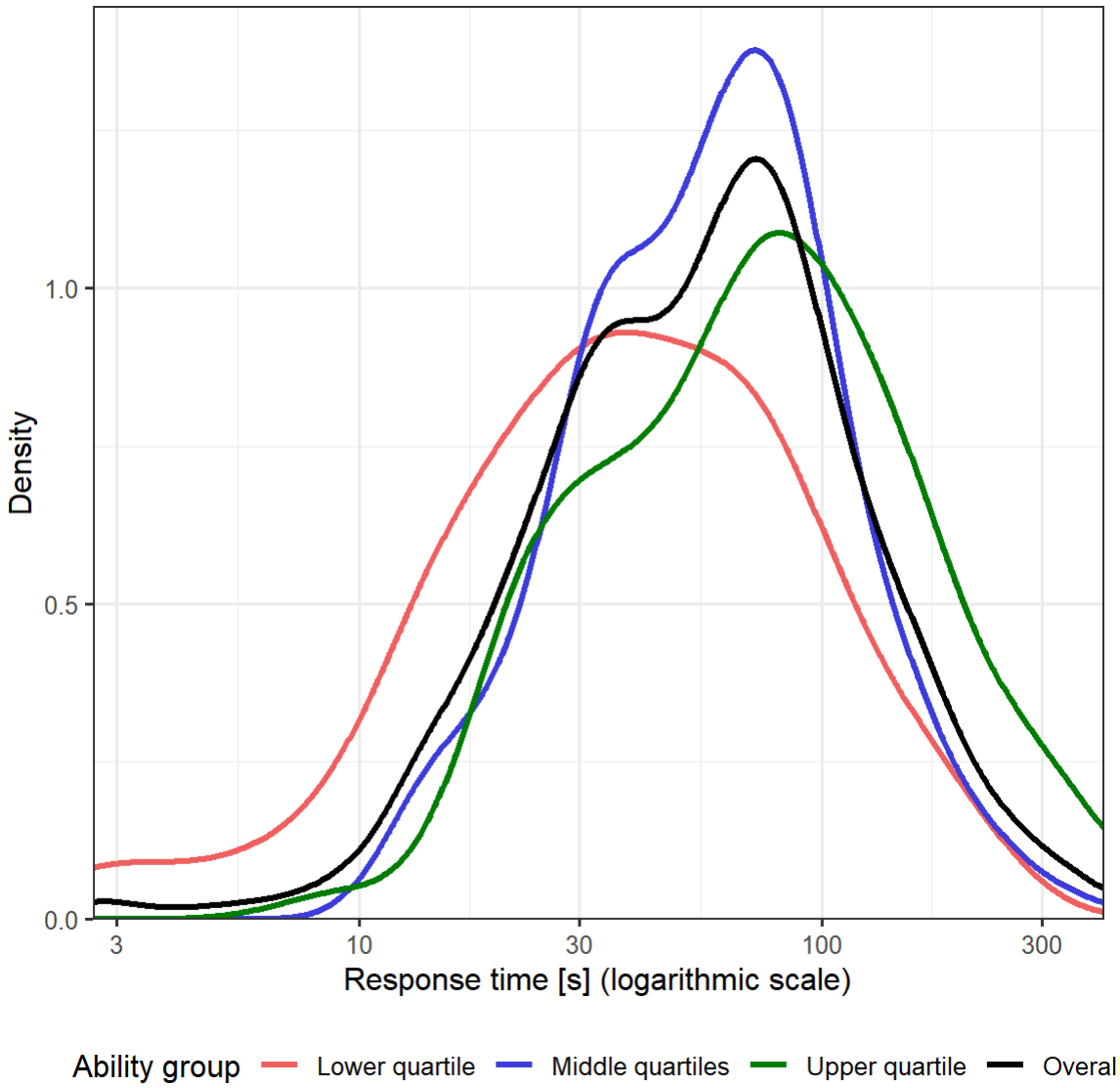
10.1.73.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.81	0.666	1.000	0.004	N/A

10.1.73.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

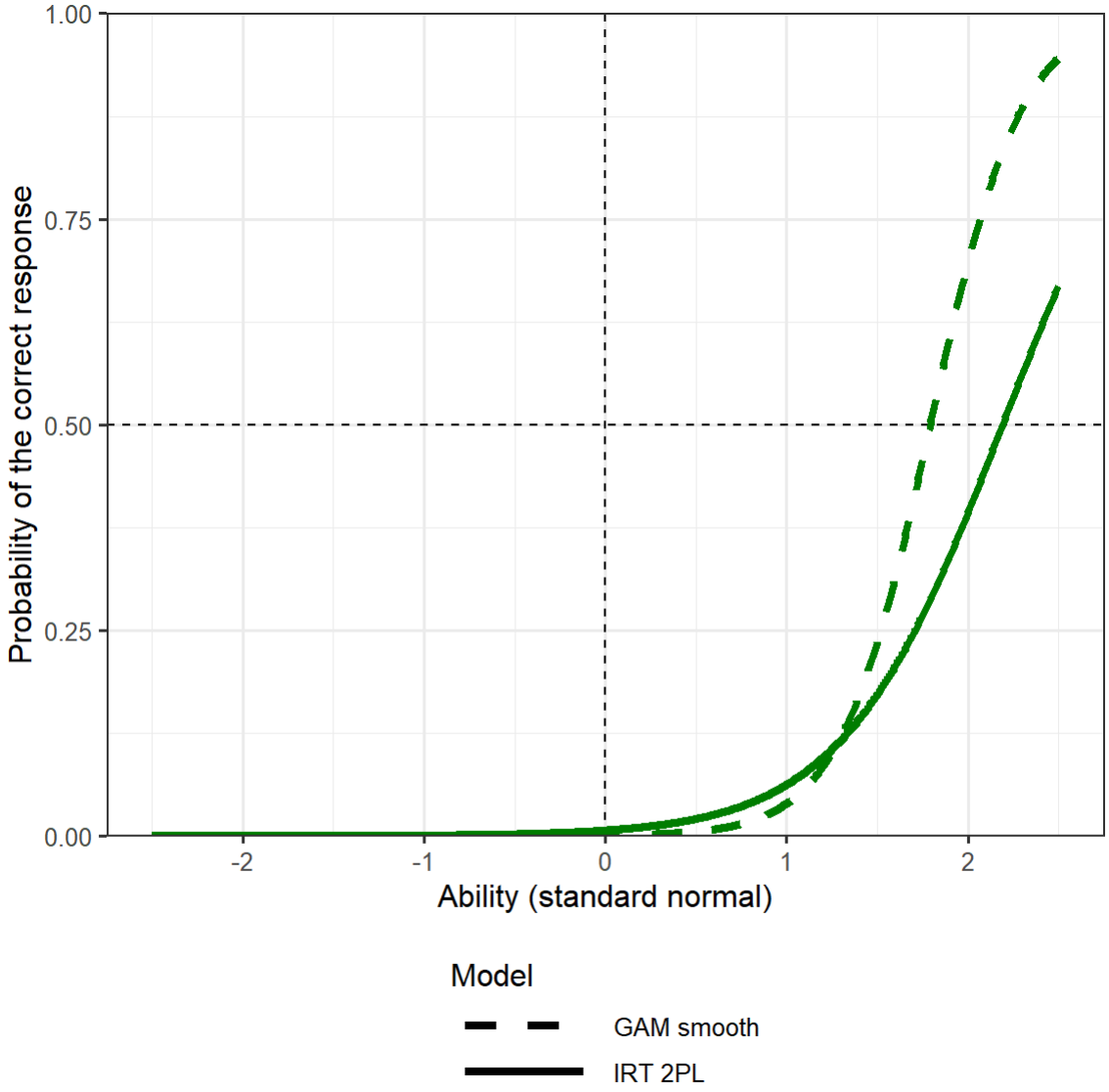
10.1.73.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	21.1	36.8	70.1	209.5	3.6	0.9
Middle quartiles	12.0	34.9	58.9	86.4	348.4	4.0	0.7
Upper quartile	9.1	37.7	73.6	123.1	406.3	4.3	0.8
Overall	2.7	32.0	58.8	89.0	406.3	4.0	0.8

10.1.74 M890004 (LLM)

10.1.74.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	3.57%	281	LLM complex setup	CR	0.31	0.17	2.27	2.19	0.00

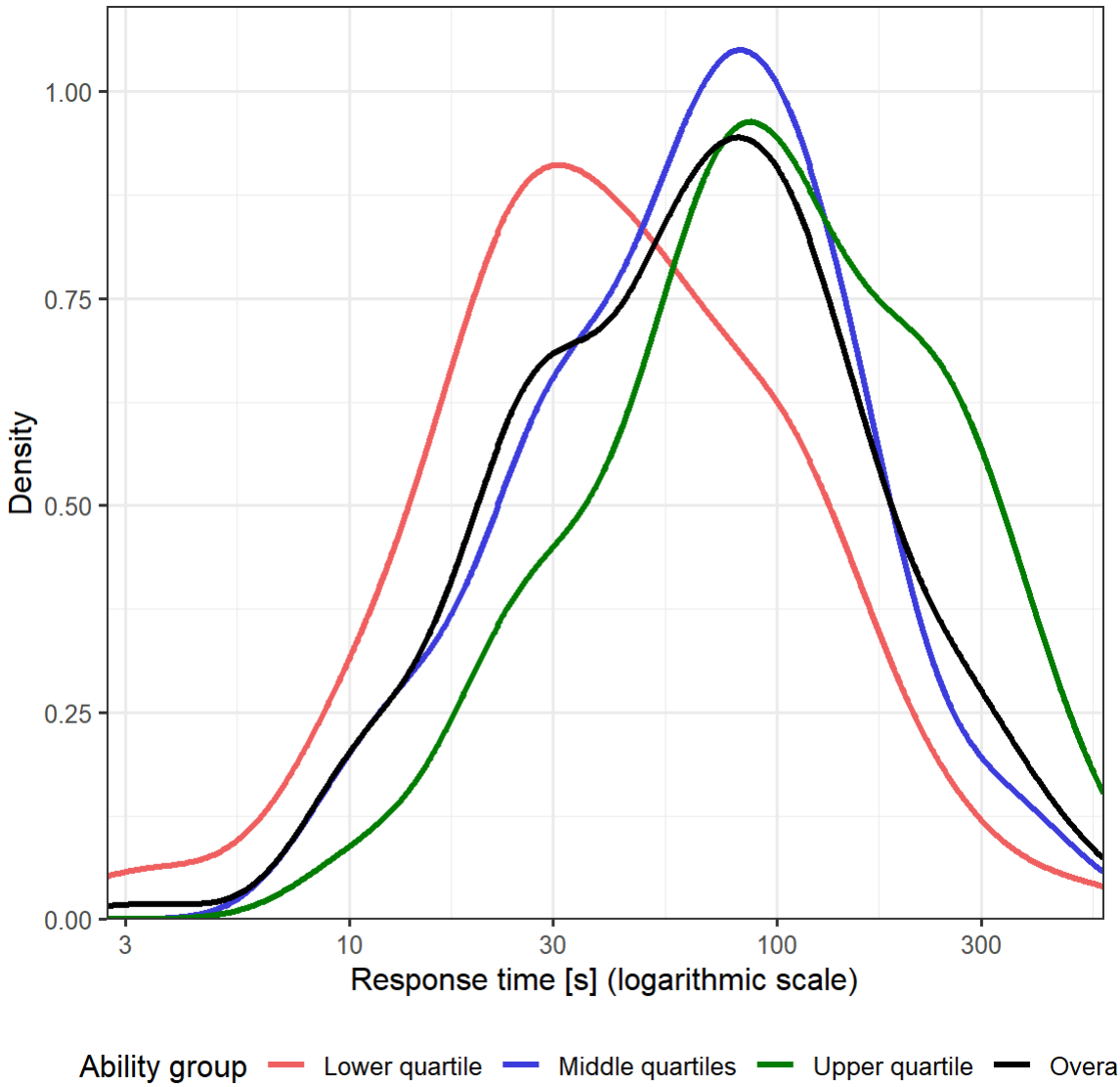
10.1.74.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.89	0.235	1.000	0.034	N/A

10.1.74.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

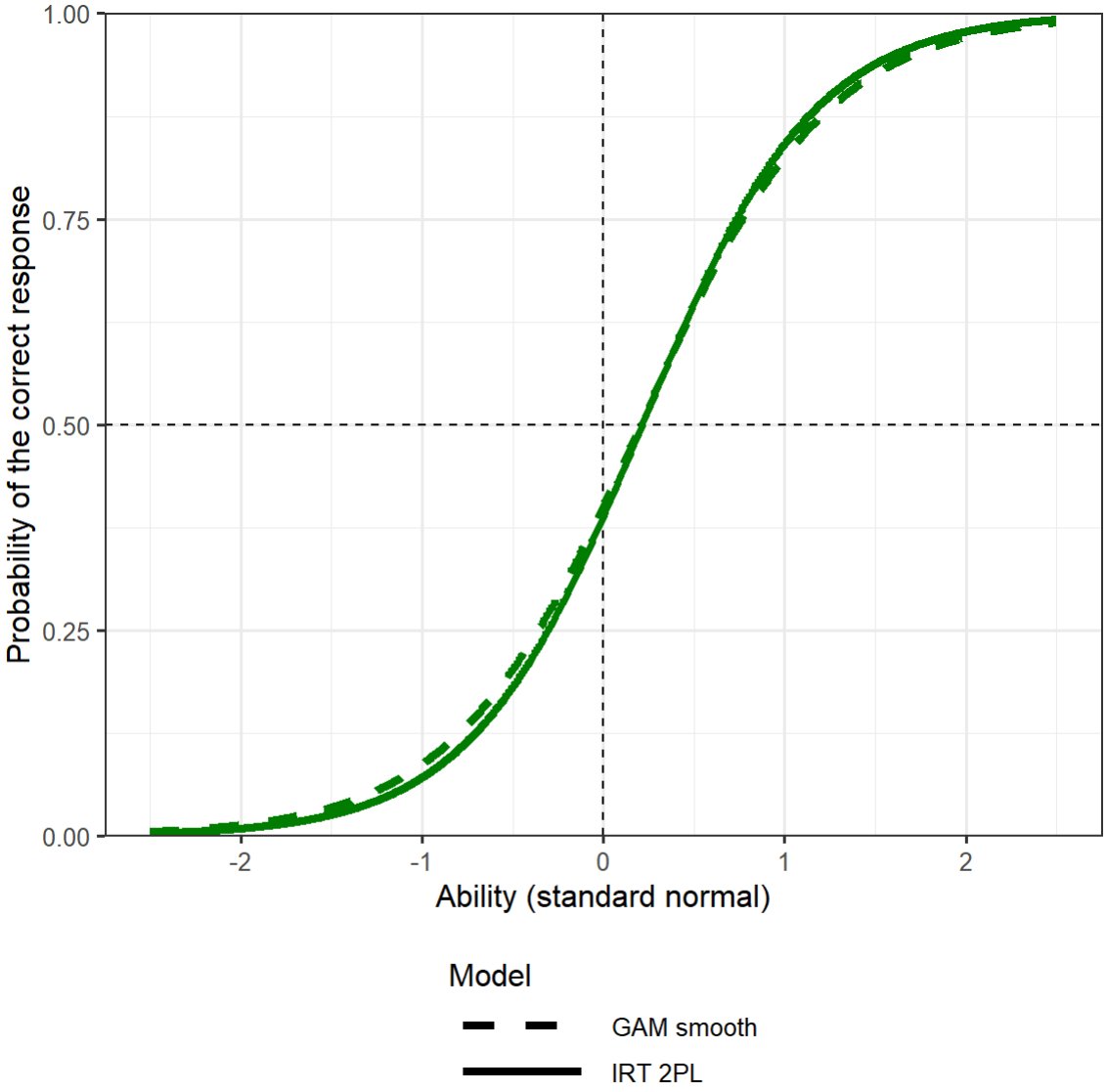
10.1.74.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	21.2	39.7	86.0	536.2	3.7	1.0
Middle quartiles	8.4	33.4	67.7	112.5	566.4	4.1	0.9
Upper quartile	10.3	55.3	92.9	191.3	578.5	4.5	0.9
Overall	2.7	30.9	67.4	121.0	578.5	4.1	1.0

10.1.75 M041059 (TIMSS)

10.1.75.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	44.97%	299	TIMSS	CR	0.60	0.62	2.12	0.21	0.029

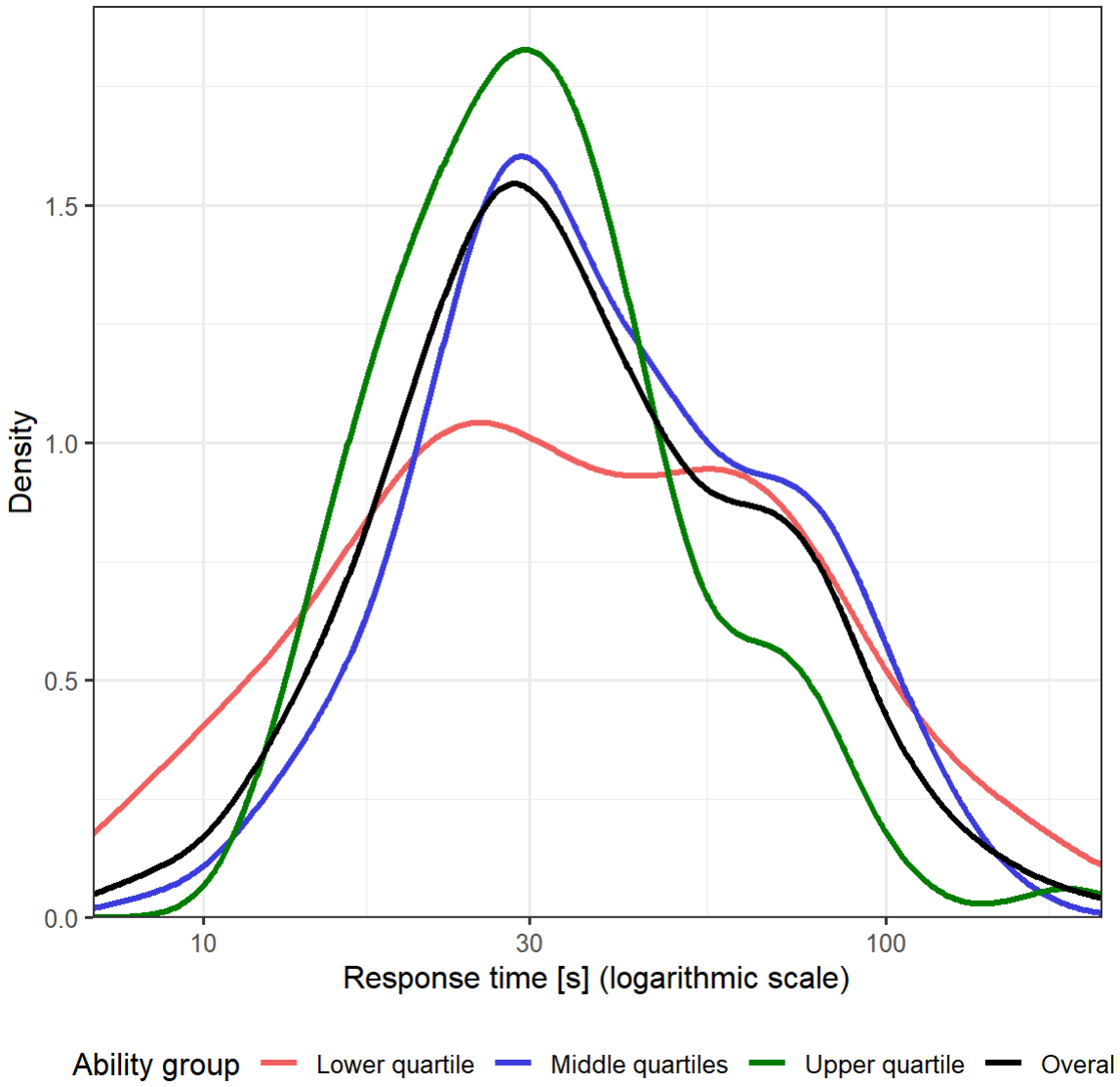
10.1.75.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
5.81	0.055	1.000	0.014	N/A

10.1.75.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

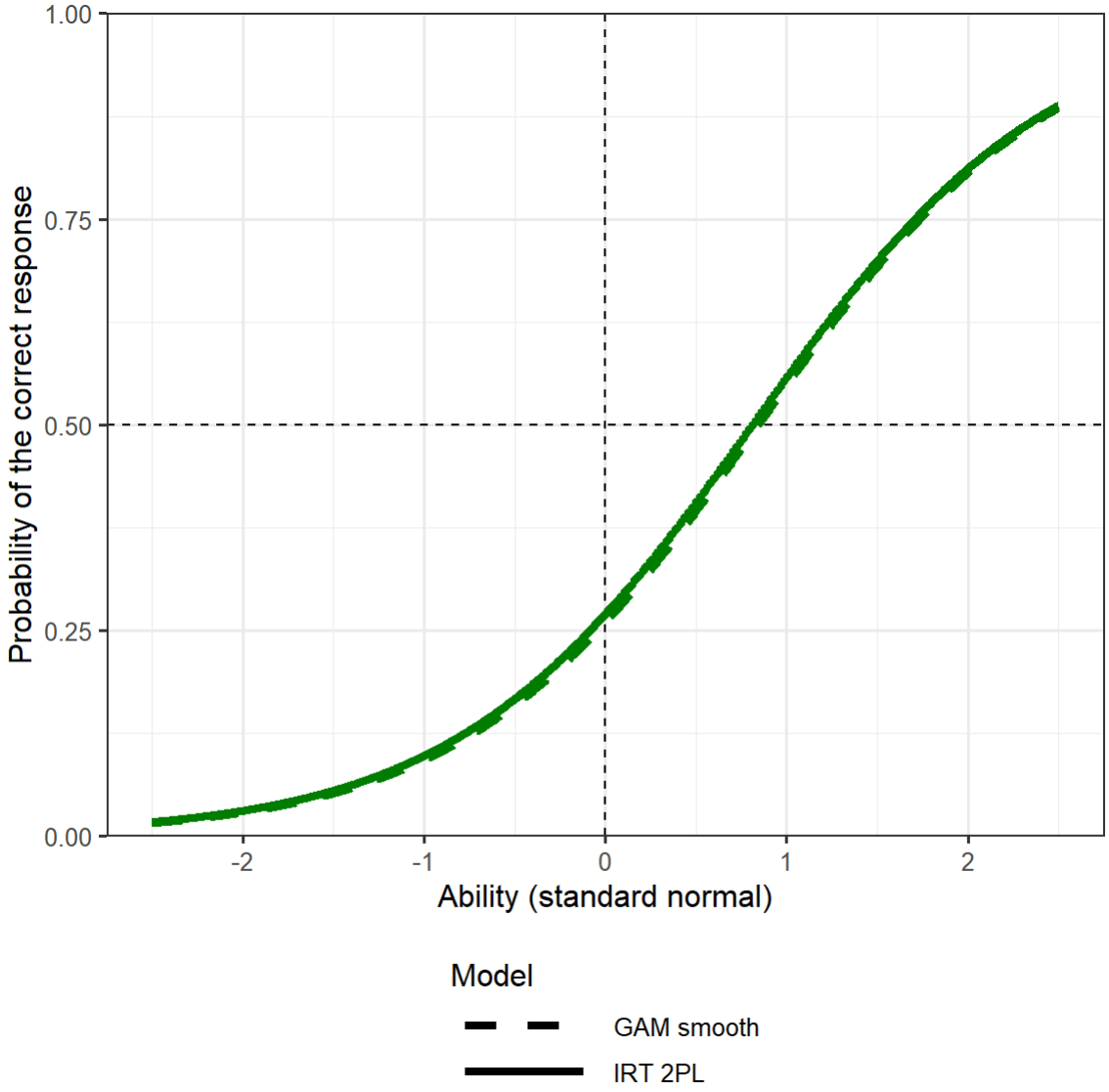
10.1.75.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	6.9	20.5	33.4	62.5	206.9	3.5	0.8
Middle quartiles	8.6	25.7	35.2	58.7	147.9	3.6	0.6
Upper quartile	13.2	21.6	30.5	41.1	185.1	3.4	0.5
Overall	6.9	23.1	32.3	55.1	206.9	3.6	0.6

10.1.76 M090057 (LLM)

10.1.76.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS generated items: Discrimination	Item difficulty	RMSE
1	31.91%	304	LLM teachers' selection	CR	0.44	0.38	1.23	0.81	0.0

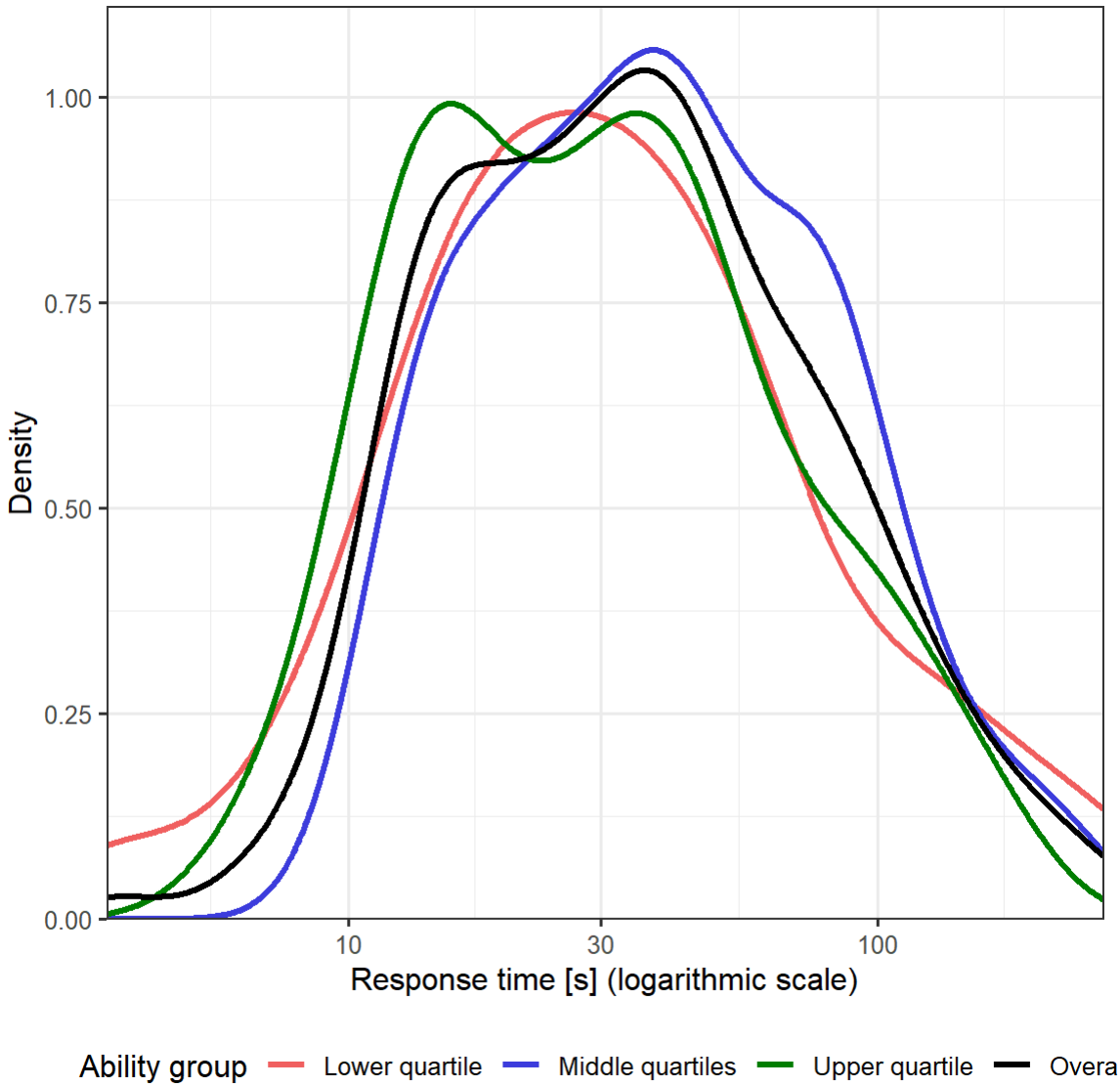
10.1.76.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.68	0.261	1.000	0.008	N/A

10.1.76.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

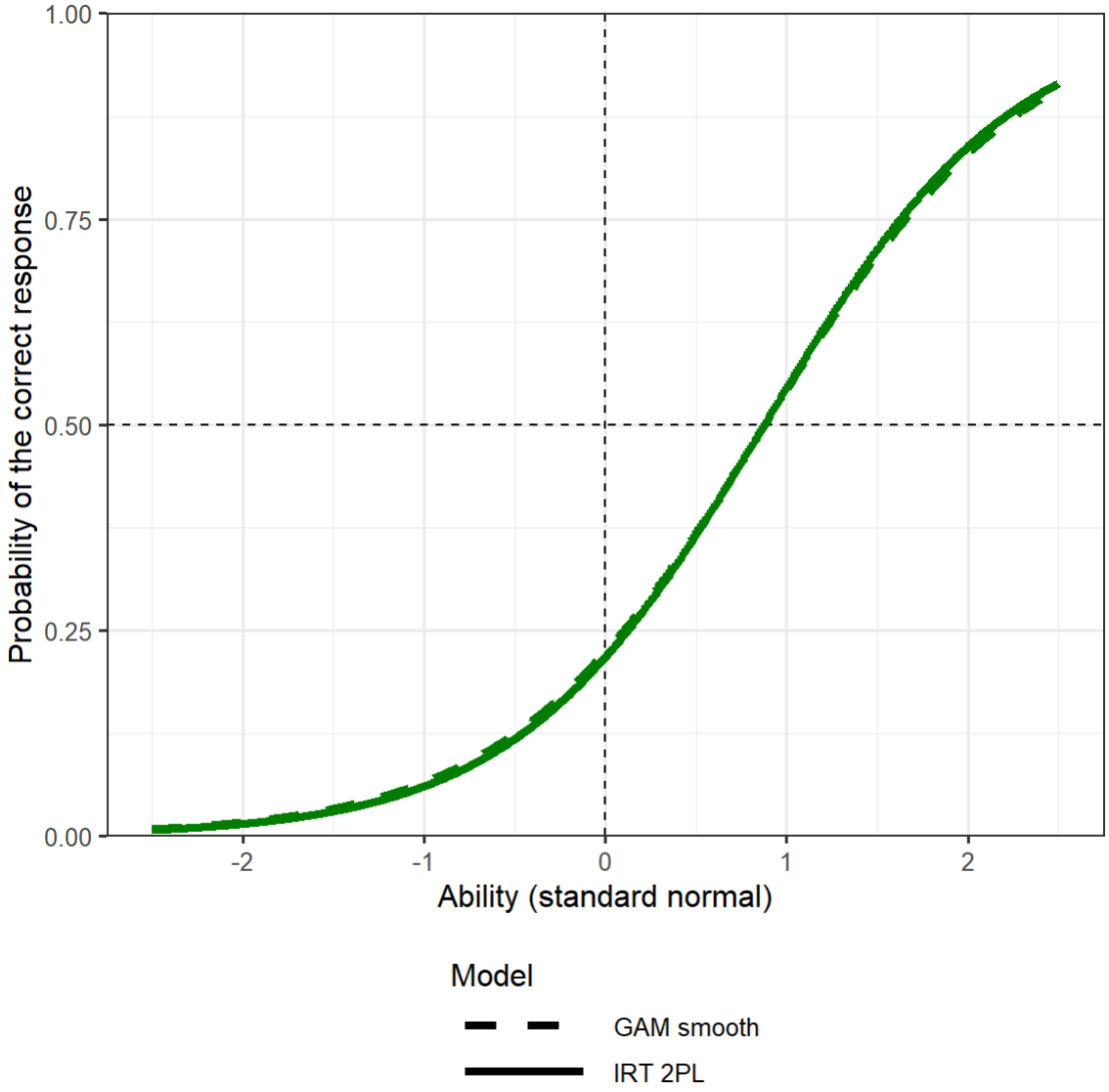
10.1.76.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	17.5	30.6	57.2	267.2	3.4	0.9
Middle quartiles	10.6	21.6	37.8	70.9	257.3	3.7	0.7
Upper quartile	6.4	14.9	29.7	46.9	160.2	3.4	0.8
Overall	3.5	18.1	33.7	59.3	267.2	3.5	0.8

10.1.77 MP71162 (TIMSS)

10.1.77.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	29.59%	294	TIMSS	CR	0.48	0.48	1.46	0.88	0.000

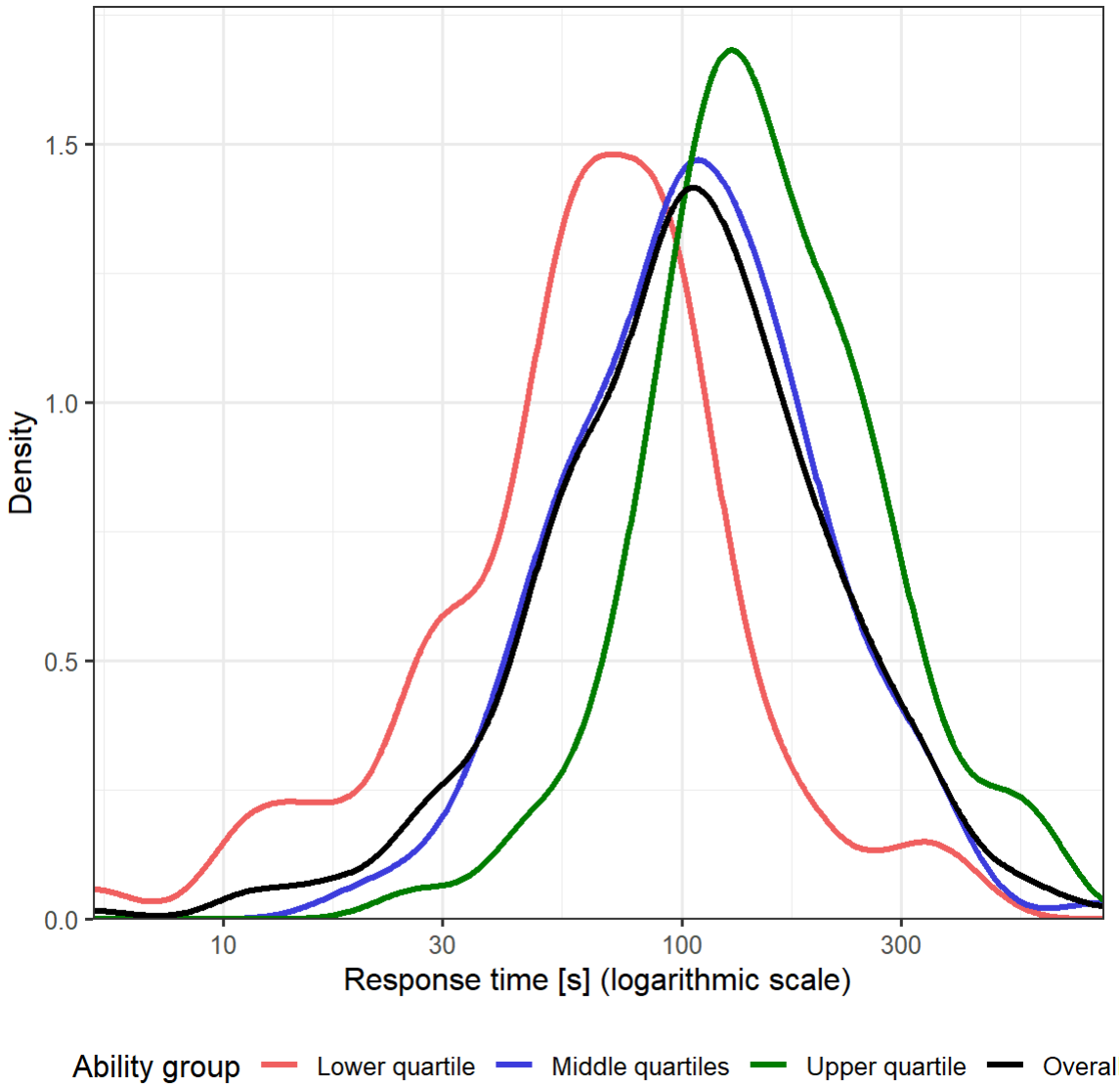
10.1.77.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.44	0.801	1.000	0.001	N/A

10.1.77.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

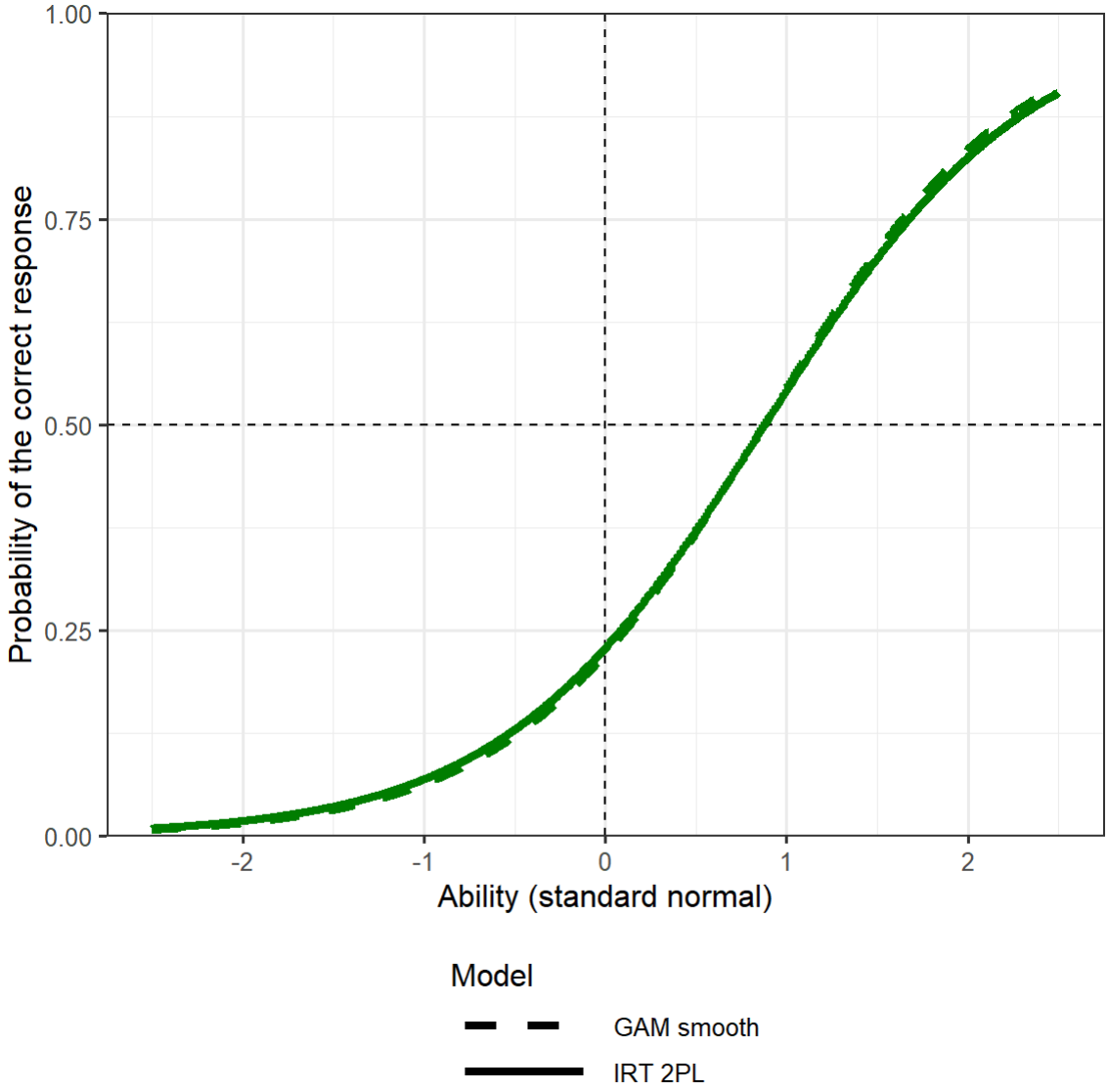
10.1.77.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.2	44.8	61.4	95.3	392.0	4.1	0.8
Middle quartiles	18.5	67.3	105.4	156.4	829.7	4.7	0.6
Upper quartile	26.4	104.9	140.7	223.7	600.1	5.0	0.6
Overall	5.2	63.7	103.6	156.9	829.7	4.6	0.7

10.1.78 M051205 (TIMSS)

10.1.78.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	27.87%	306	TIMSS	CR	0.47	0.50	1.38	0.88	0.000

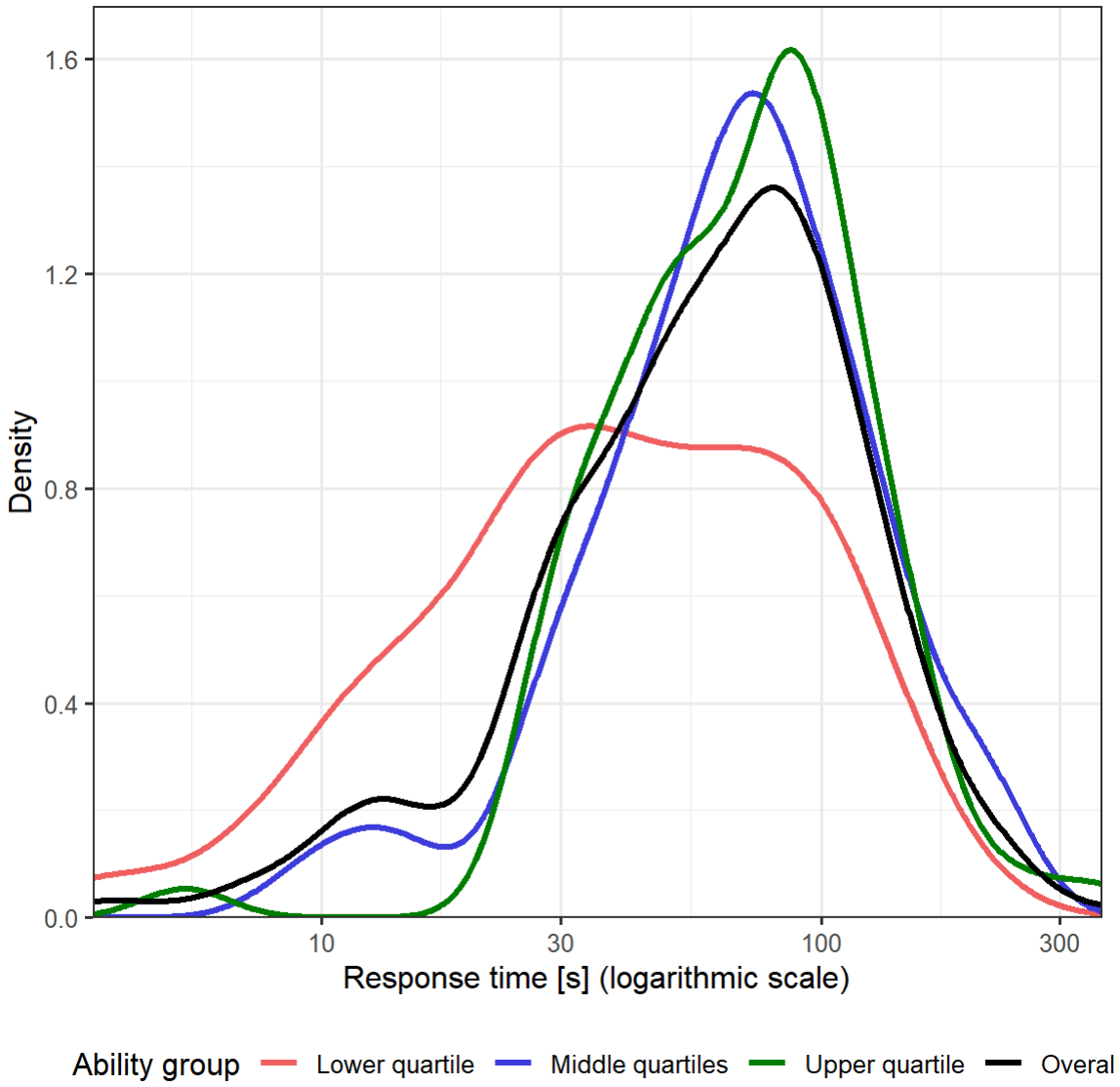
10.1.78.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
4.50	0.105	1.000	0.013	N/A

10.1.78.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

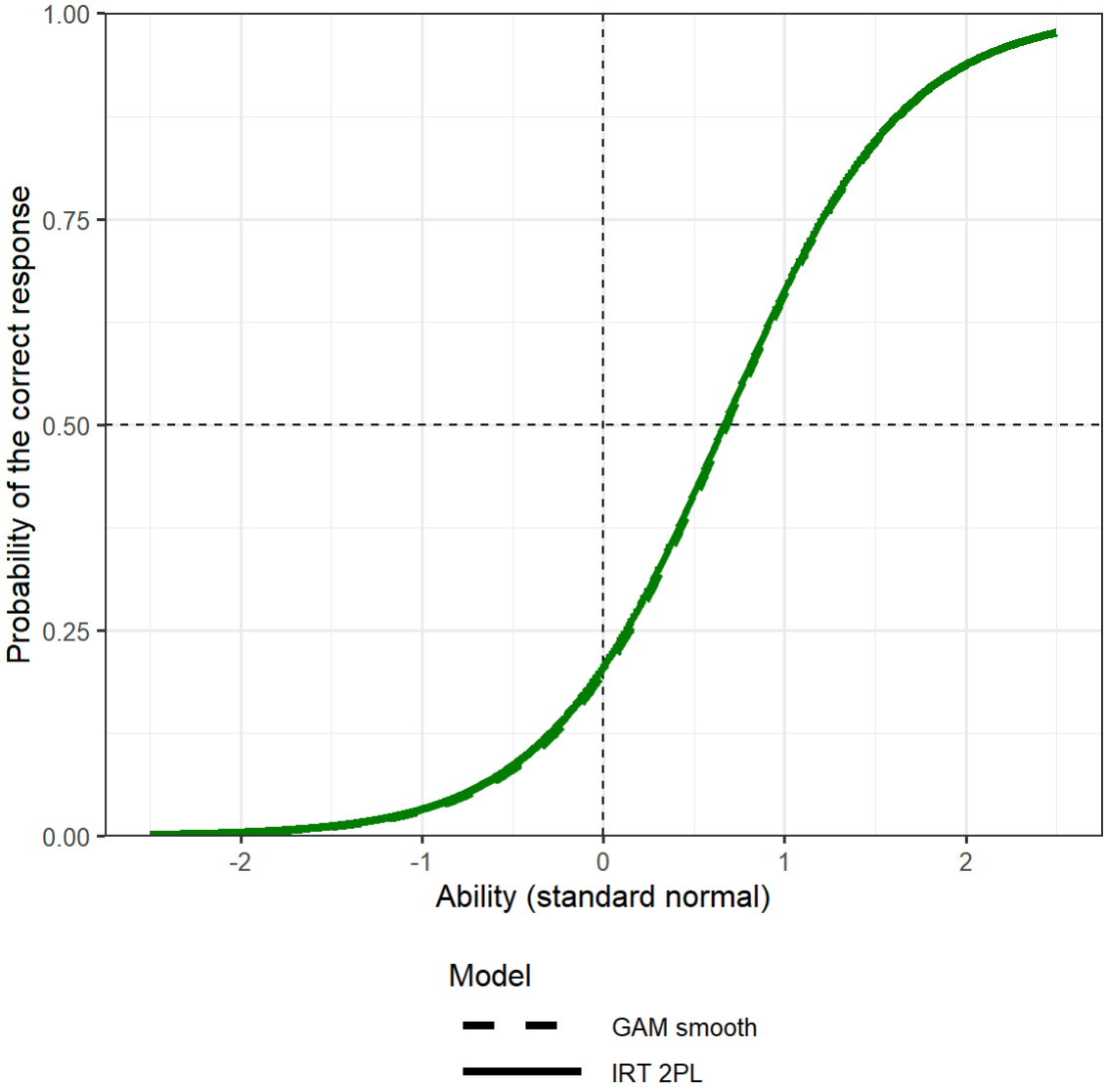
10.1.78.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	22.7	40.6	78.8	186.4	3.6	0.9
Middle quartiles	8.2	44.9	70.9	100.0	250.8	4.2	0.7
Upper quartile	5.3	45.7	73.6	96.0	362.6	4.2	0.6
Overall	3.5	36.3	63.4	95.9	362.6	4.0	0.8

10.1.79 M890003 (LLM)

10.1.79.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	30.00%	281	LLM complex setup	CR	0.56	0.36	2.03	0.67	0.01

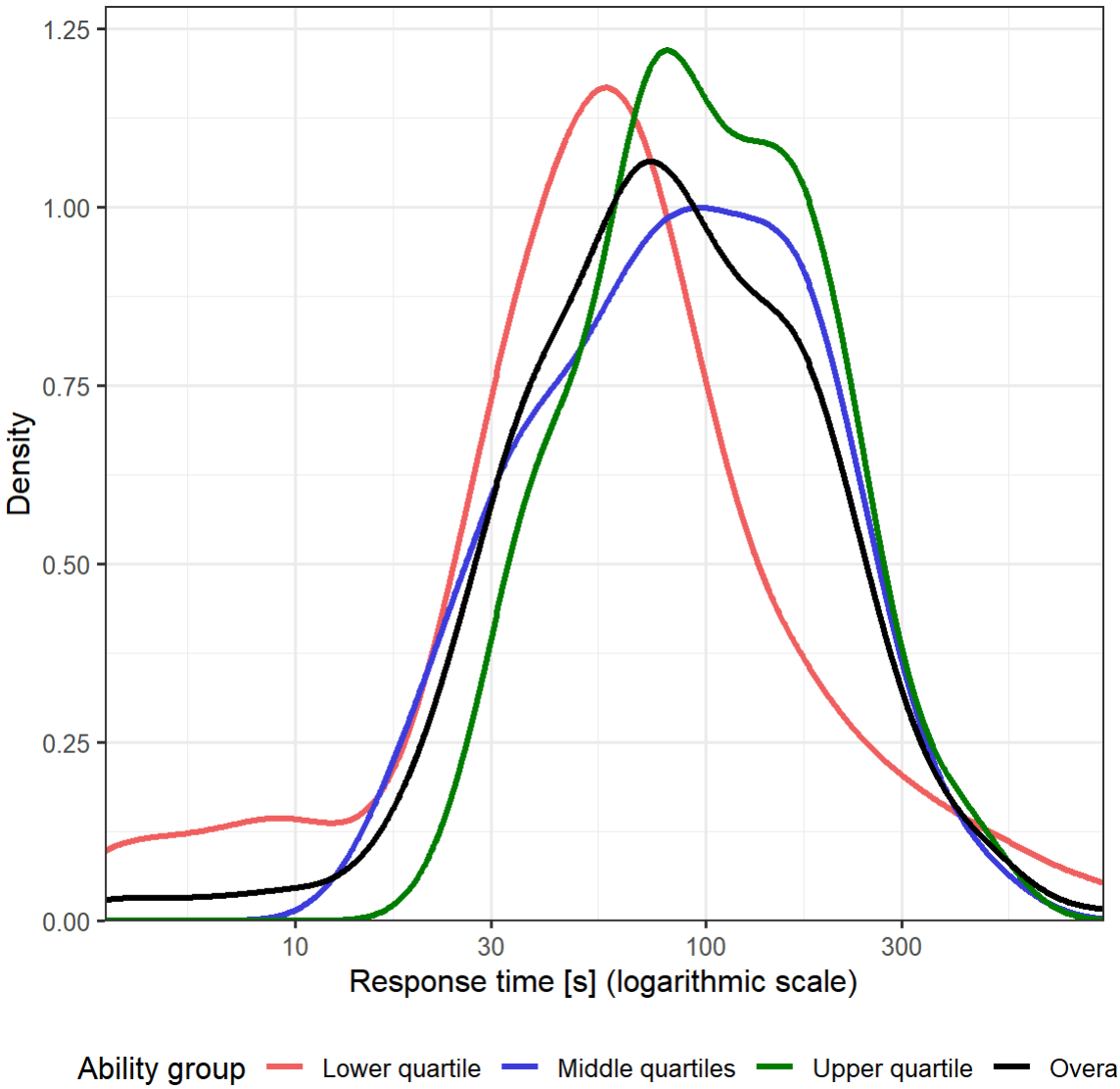
10.1.79.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.49	0.175	1.000	0.010	N/A

10.1.79.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

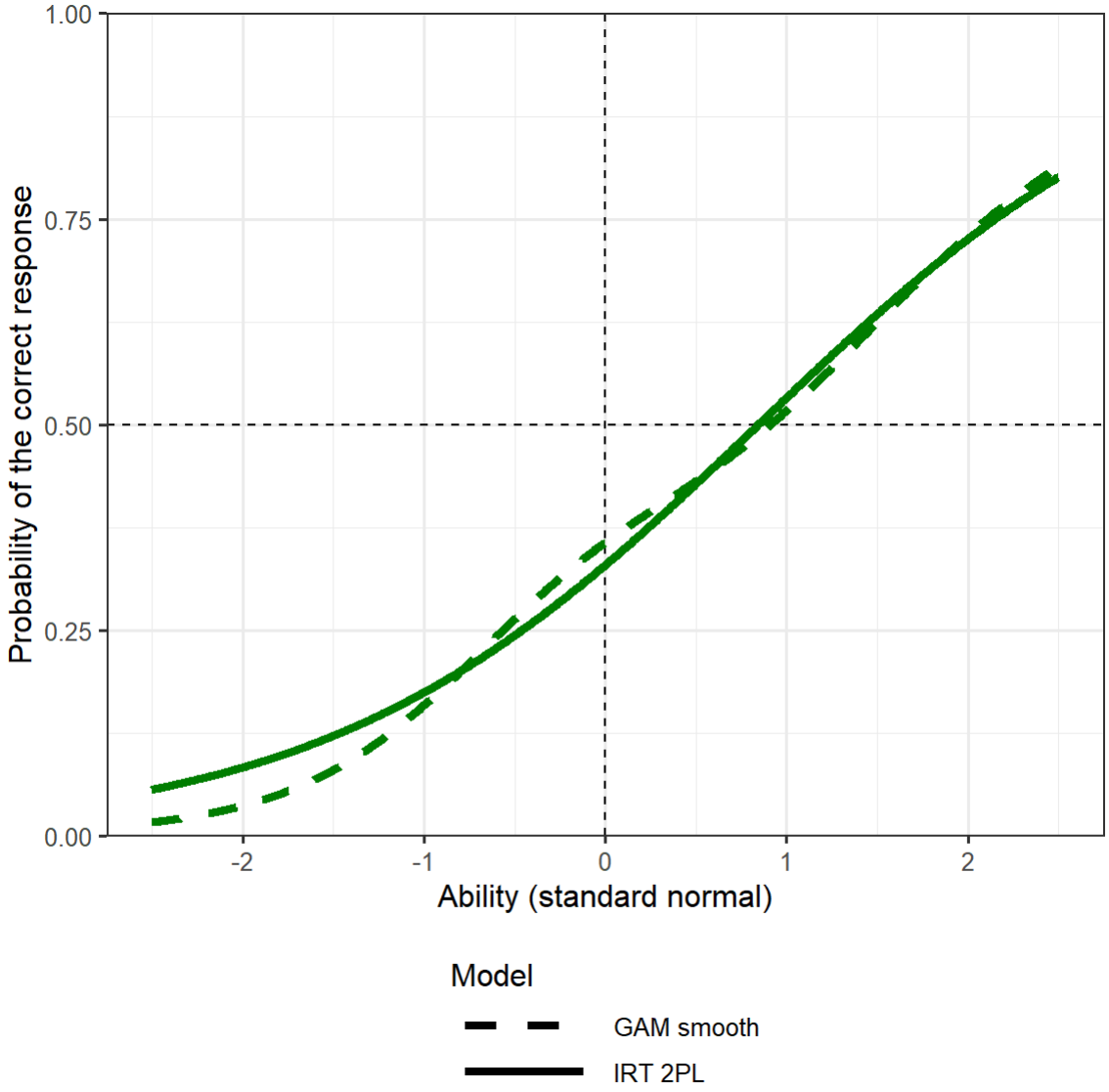
10.1.79.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.4	33.8	57.8	98.2	928.3	4.0	1.0
Middle quartiles	15.6	45.0	86.8	158.4	521.3	4.4	0.8
Upper quartile	25.9	66.5	96.0	157.8	479.6	4.6	0.7
Overall	3.4	43.8	78.3	152.9	928.3	4.4	0.9

10.1.80 M100024 (LLM)

10.1.80.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	35.42%	319	LLM simple setup	CR	0.37	0.20	0.84	0.85	0.021

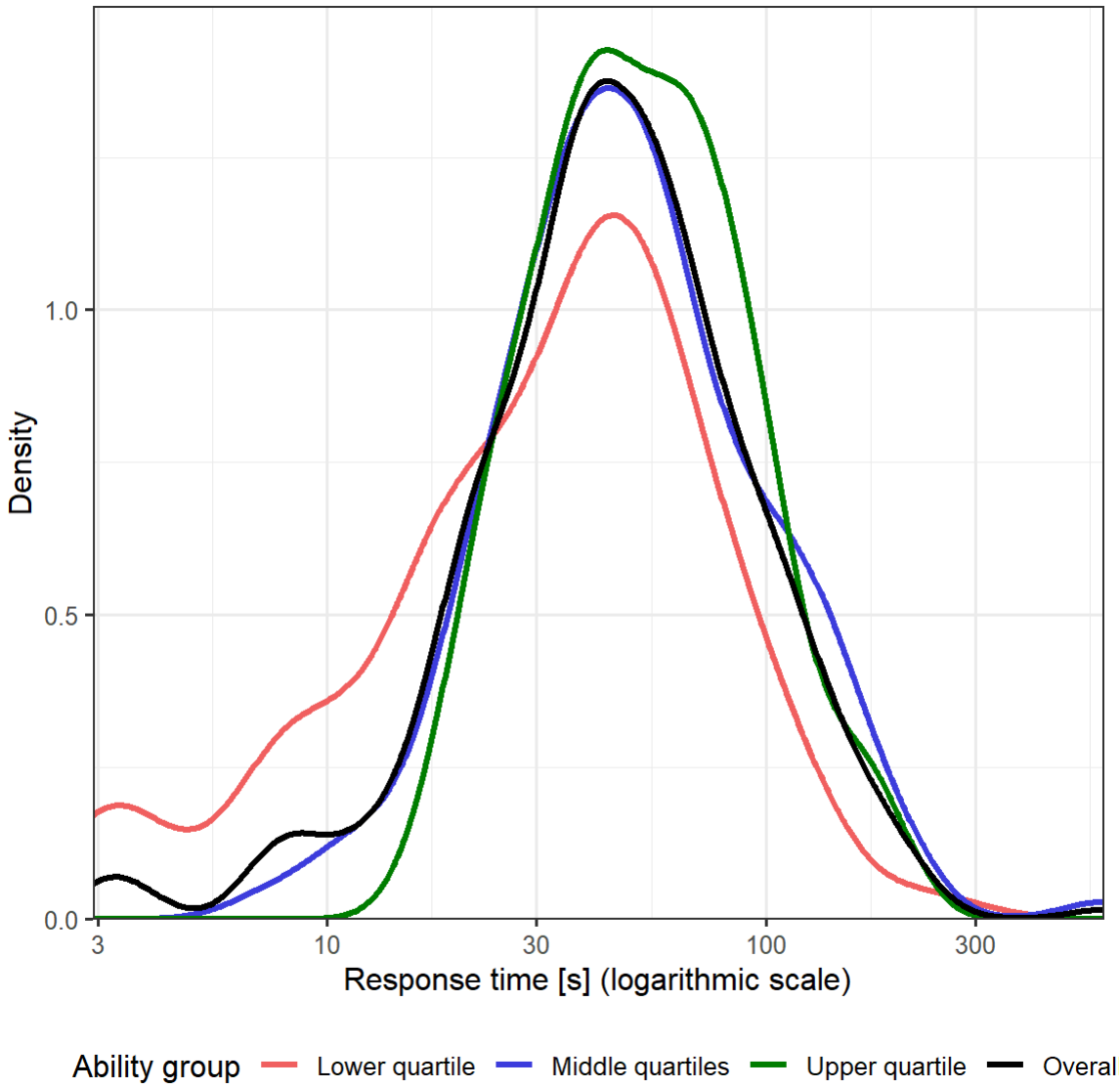
10.1.80.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.26	0.880	1.000	0.001	N/A

10.1.80.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

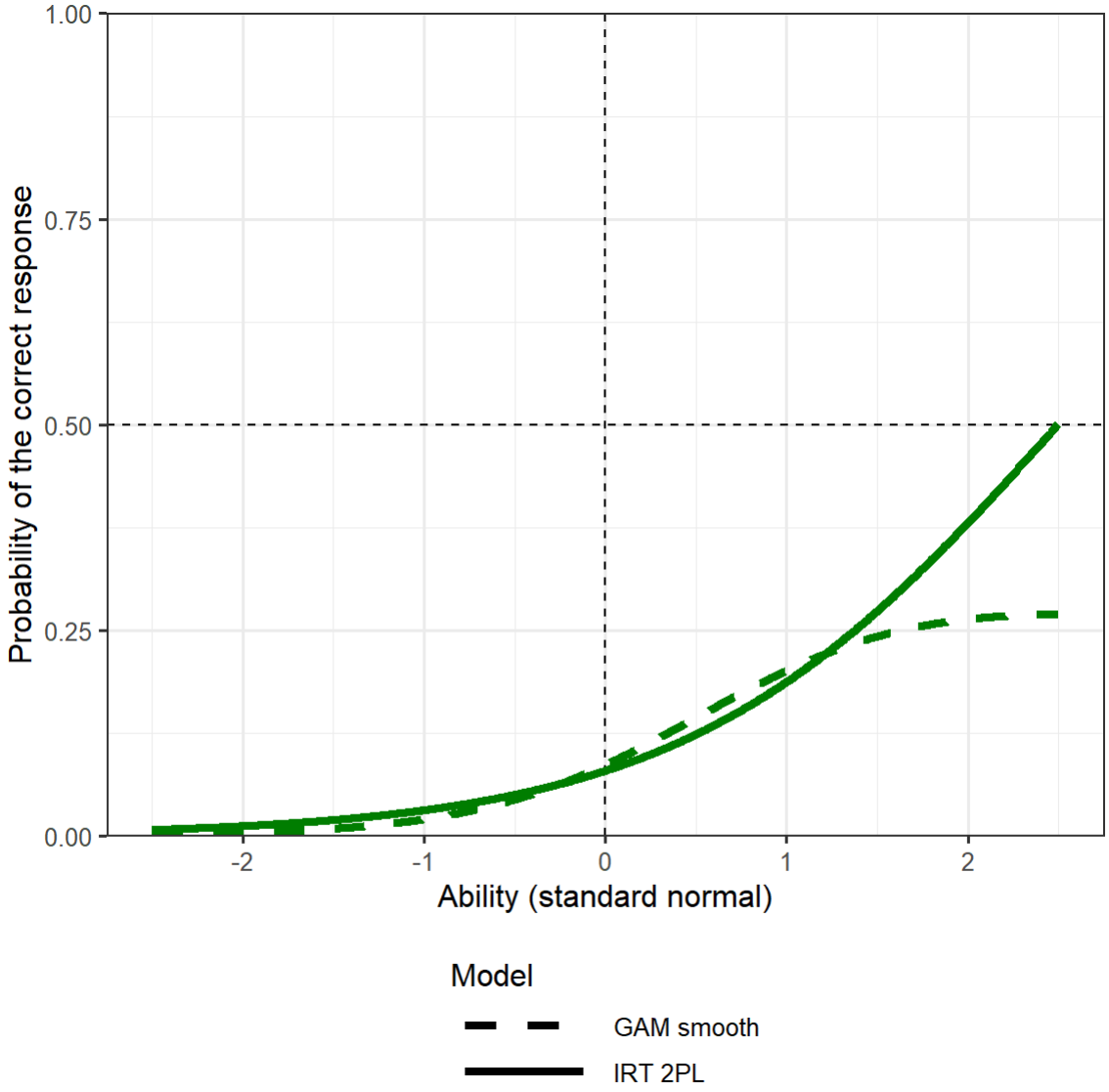
10.1.80.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.9	19.0	36.2	54.6	234.8	3.4	0.9
Middle quartiles	7.1	31.1	48.5	75.9	587.0	3.9	0.7
Upper quartile	17.4	35.2	49.9	75.1	185.6	3.9	0.6
Overall	2.9	28.5	45.0	69.5	587.0	3.8	0.8

10.1.81 M100028 (LLM)

10.1.81.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS a generated items: Discrimination	IRT model including both TIMSS a generated items: Item difficulty	IRT model including both TIMSS a generated items: RMSEA
1	10.86%	304	LLM simple setup	CR	0.24	0.16	0.98	2.50	0.000

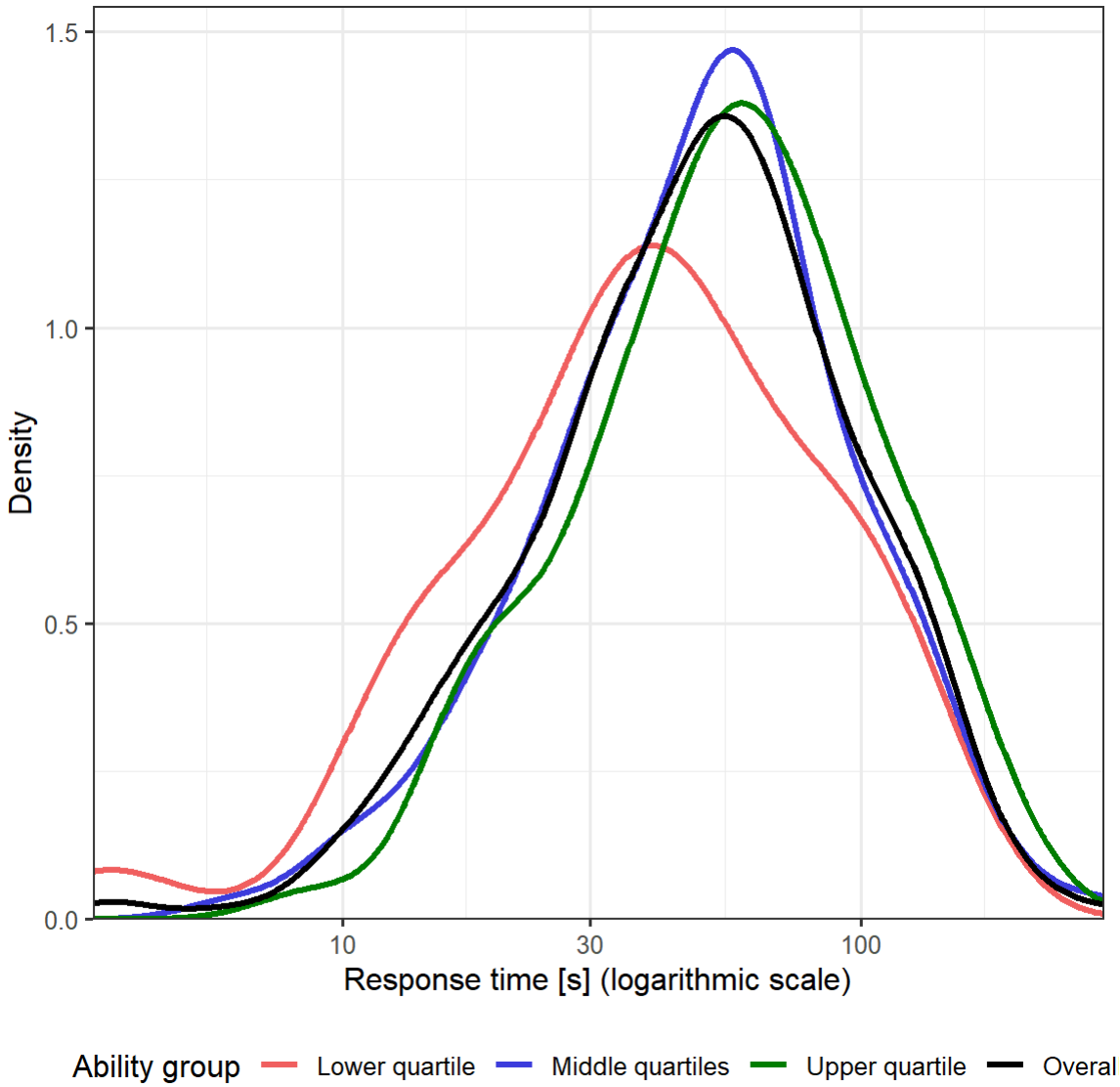
10.1.81.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.25	0.324	1.000	0.011	N/A

10.1.81.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

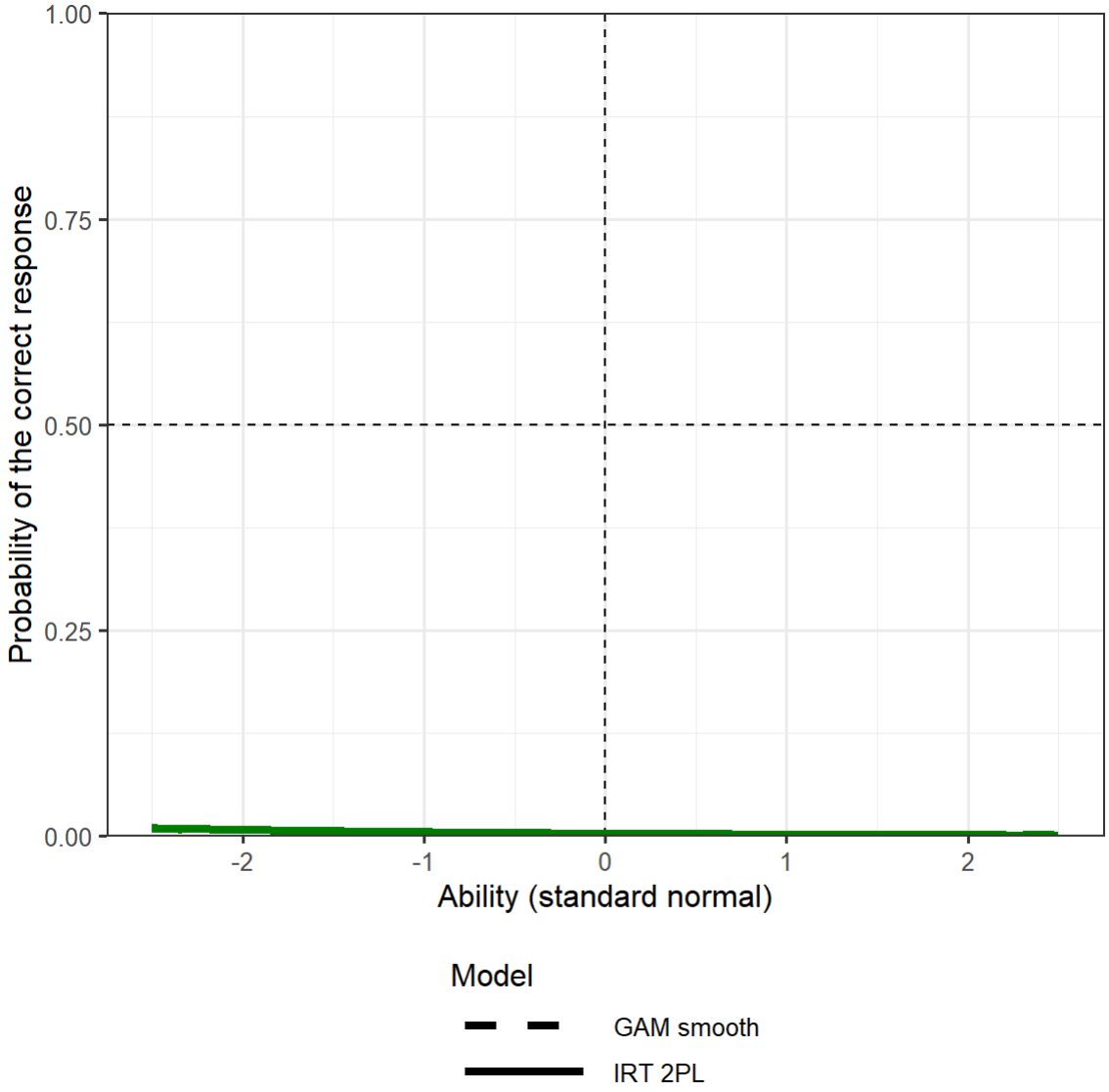
10.1.81.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.3	23.1	41.8	67.9	144.0	3.6	0.8
Middle quartiles	6.1	31.6	51.6	72.7	293.8	3.9	0.7
Upper quartile	8.7	35.9	56.0	84.4	217.0	4.0	0.7
Overall	3.3	31.0	50.7	75.4	293.8	3.9	0.7

10.1.82 M090048 (LLM)

10.1.82.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSI
1	0.34%	298	LLM teachers' selection	CR	−0.03	−0.04	−0.47	−12.40	0.0

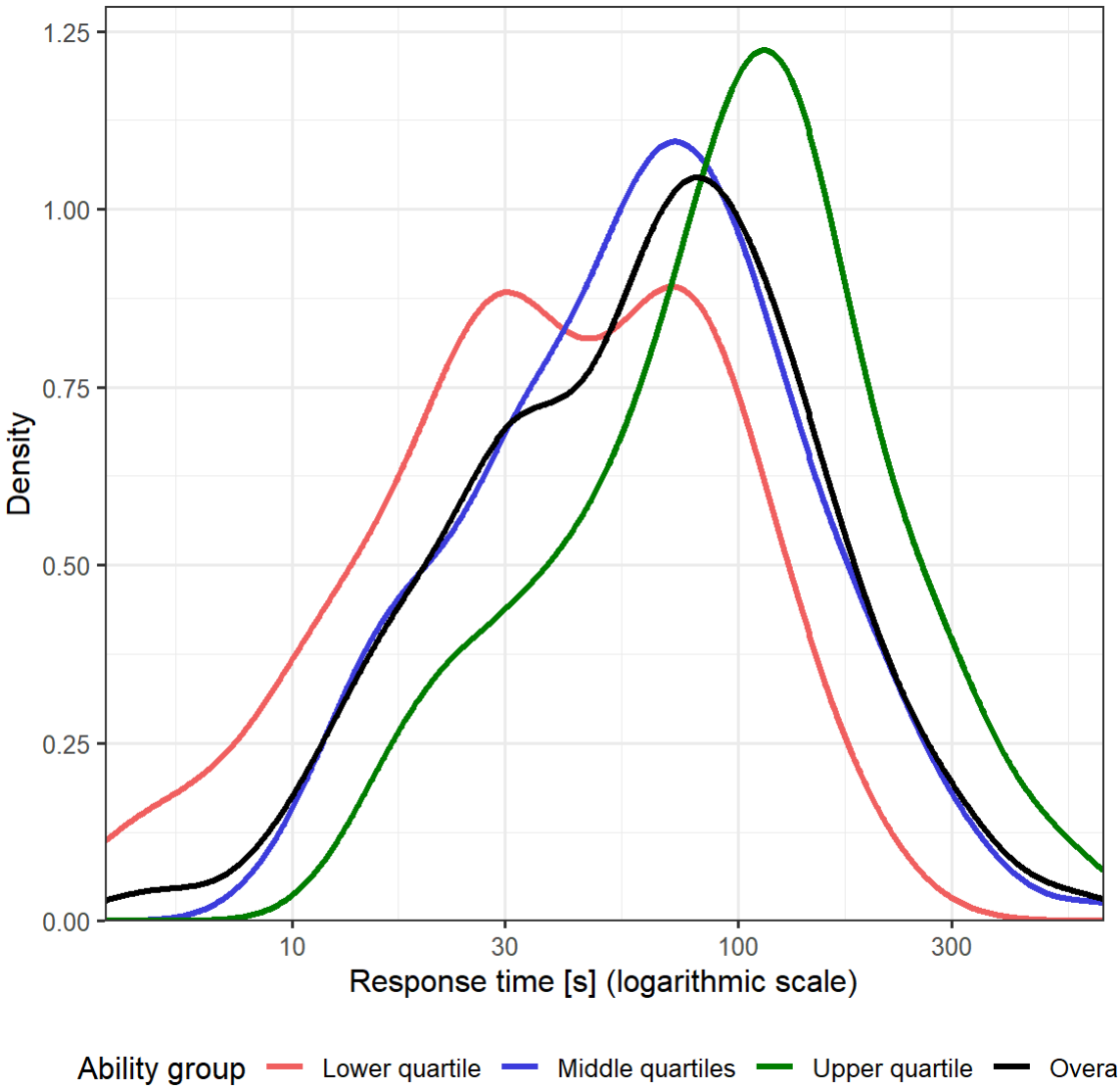
10.1.82.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.43	0.489	1.000	0.107	N/A

10.1.82.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

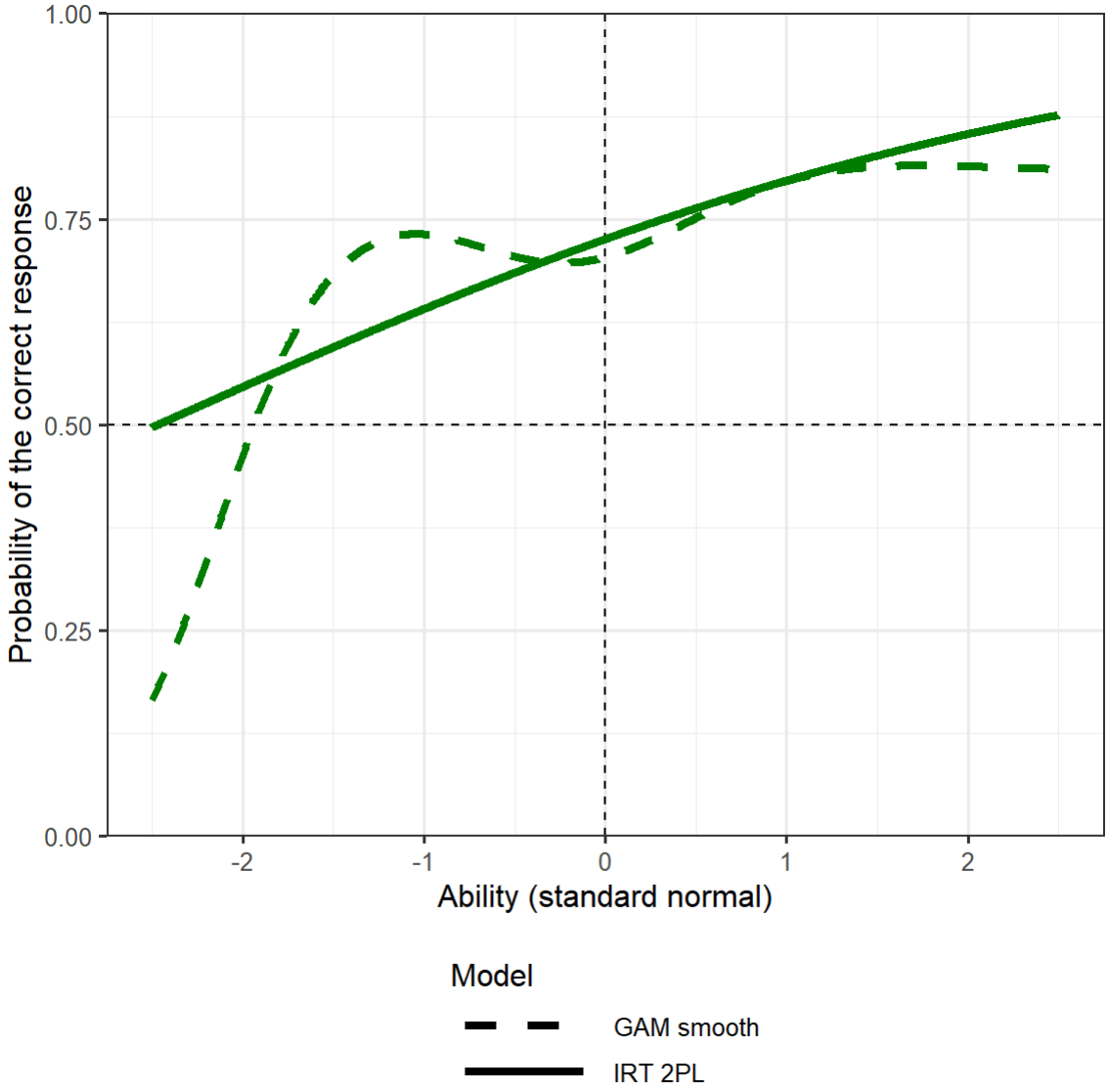
10.1.82.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.8	20.4	36.0	75.6	195.2	3.6	0.9
Middle quartiles	9.6	32.7	63.6	105.8	656.0	4.1	0.8
Upper quartile	15.4	54.6	102.7	153.7	557.4	4.5	0.8
Overall	3.8	31.6	65.4	110.0	656.0	4.1	0.9

10.1.83 M890002 (LLM)

10.1.83.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS generated items: Discrimination	IRT model including both TIMSS generated items: Item difficulty	IRT model including both TIMSS generated items: RMSE
1	71.79%	281	LLM complex setup	CR	0.19	0.08	0.39	−2.47	0.00

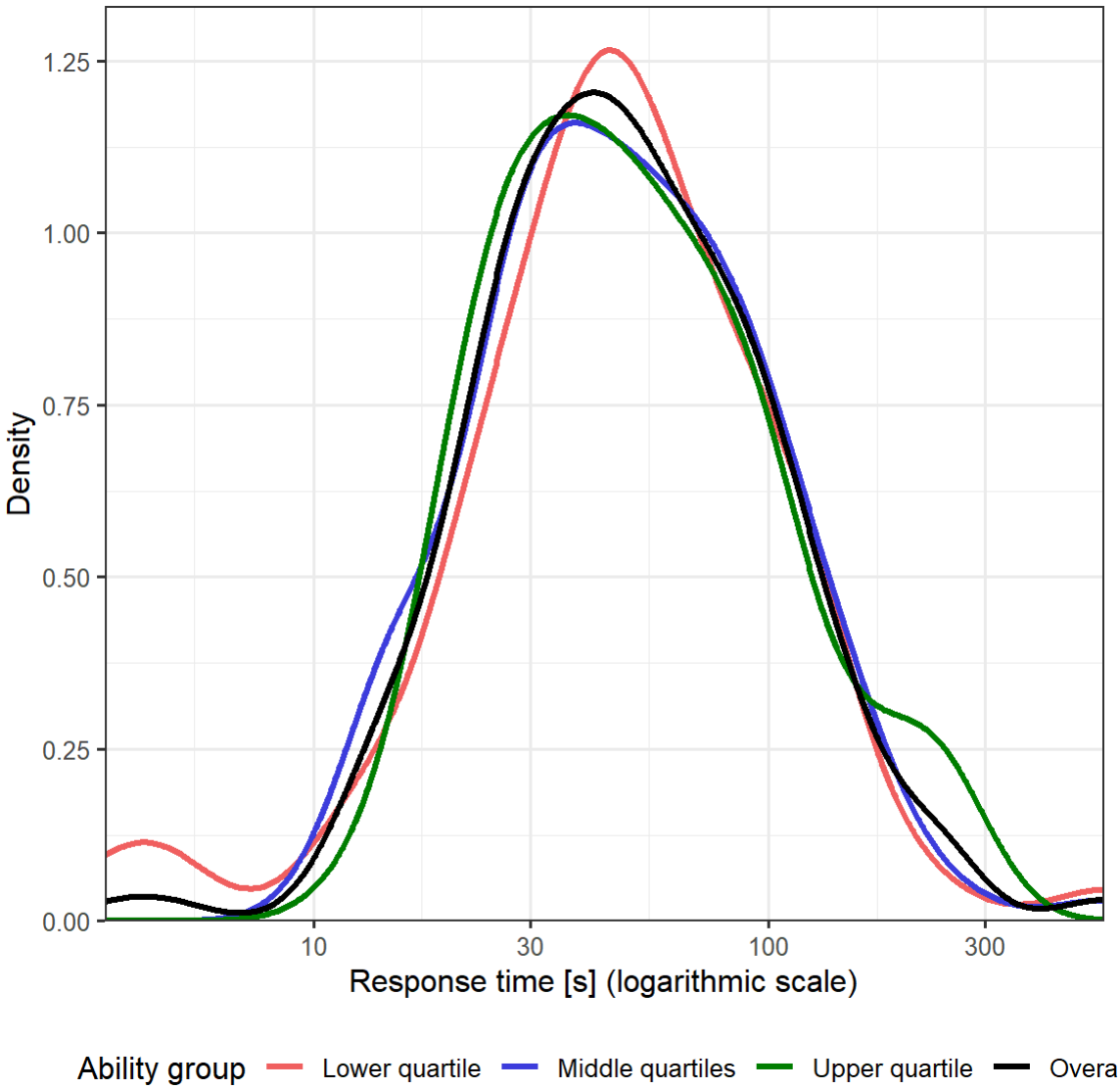
10.1.83.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.20	0.550	1.000	0.004	N/A

10.1.83.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

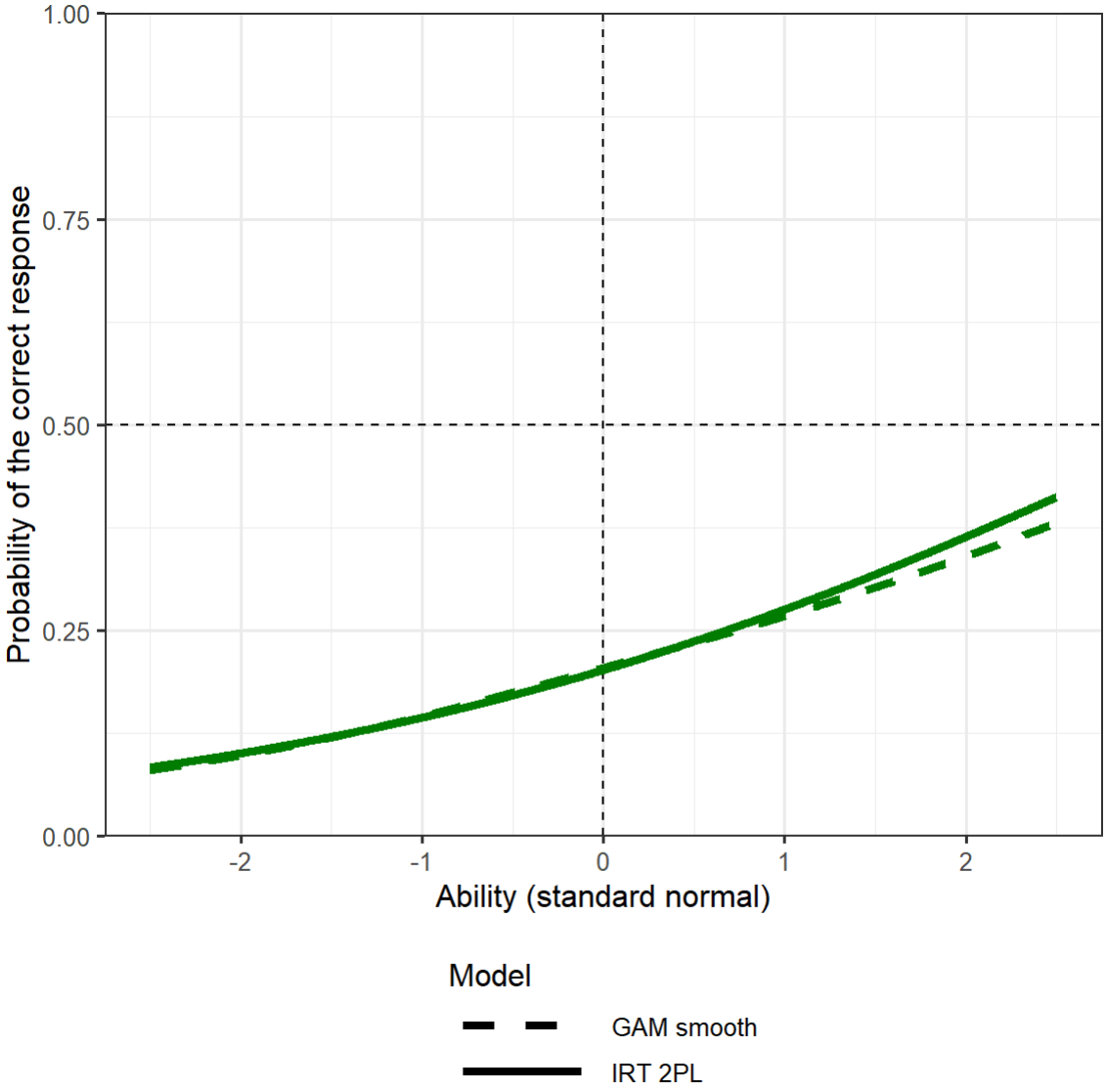
10.1.83.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	29.4	45.4	76.5	544.9	3.8	0.8
Middle quartiles	10.8	29.0	48.2	80.7	539.1	3.9	0.7
Upper quartile	13.1	29.2	47.4	80.9	249.9	3.9	0.7
Overall	3.5	29.2	46.3	79.8	544.9	3.9	0.7

10.1.84 M090007 (LLM)

10.1.84.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS generated items: Discrimination	Item difficulty	RMSE
1	21.05%	304	LLM teachers' selection	CR	0.15	0.11	0.41	3.37	0.03

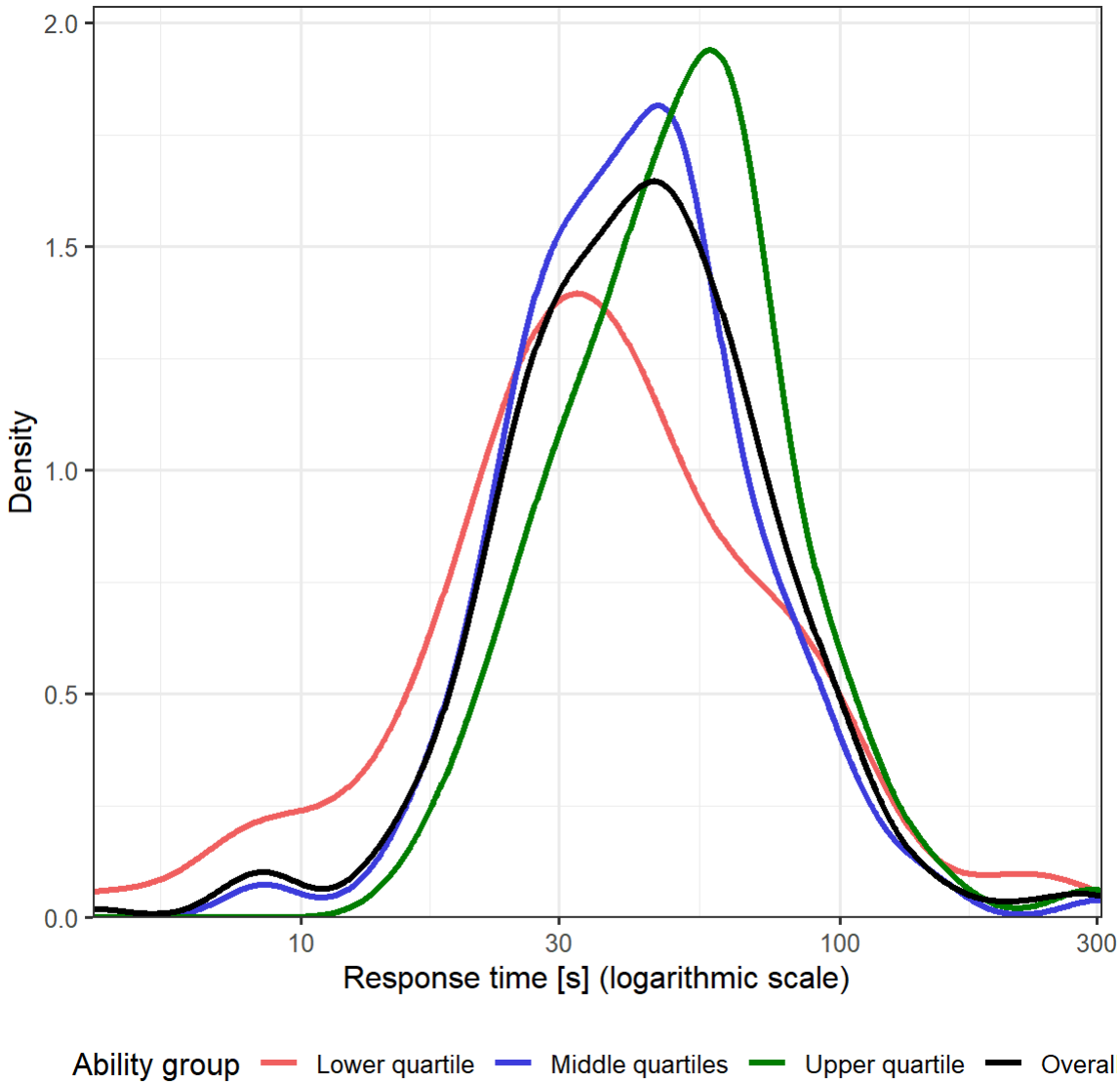
10.1.84.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
5.06	0.080	1.000	0.018	N/A

10.1.84.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

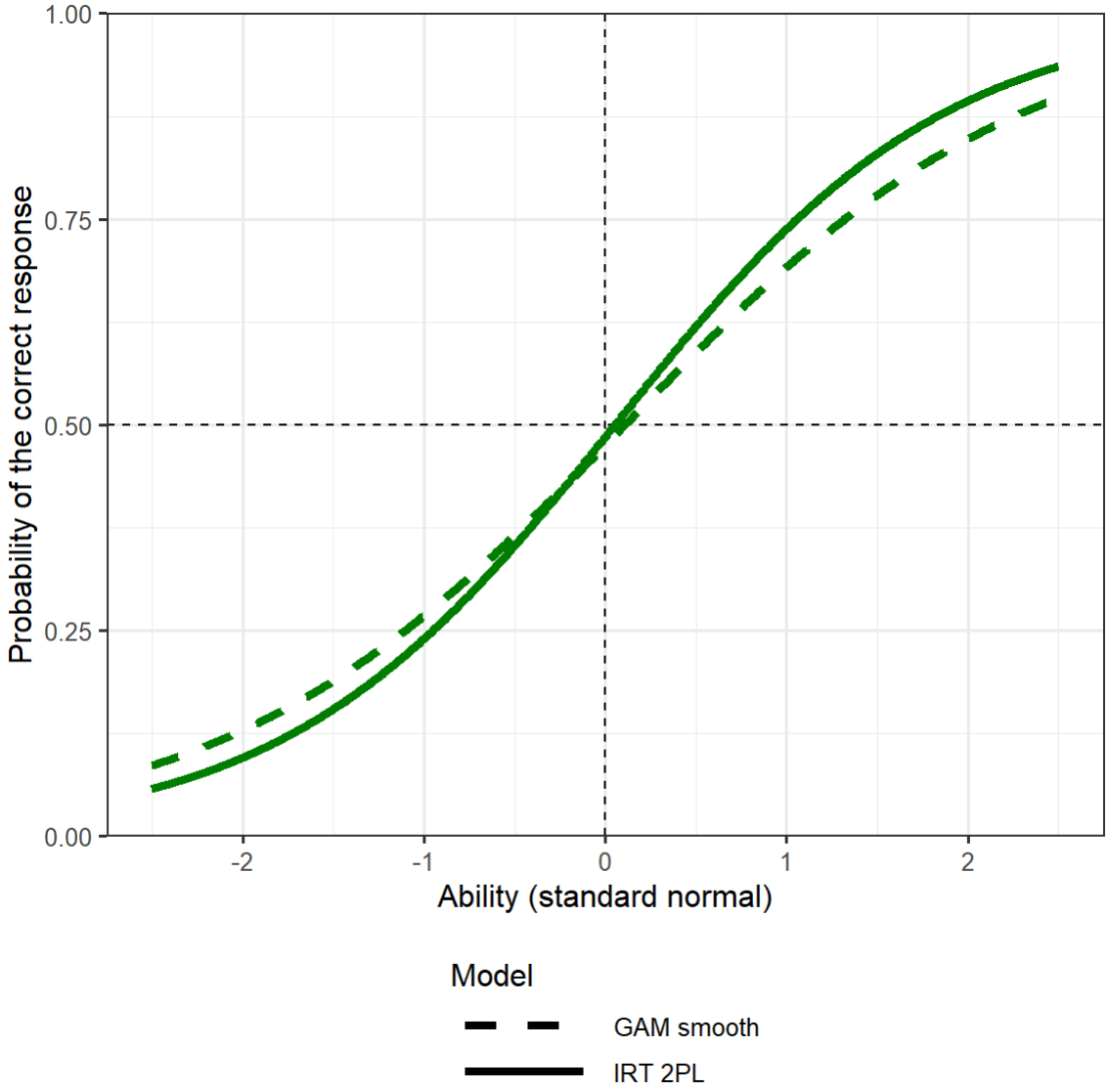
10.1.84.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.1	23.6	34.9	58.7	254.3	3.6	0.7
Middle quartiles	8.0	28.6	41.1	54.1	305.4	3.7	0.5
Upper quartile	16.1	35.6	50.1	66.8	294.8	3.9	0.5
Overall	4.1	28.9	43.0	60.6	305.4	3.7	0.6

10.1.85 M051061 (TIMSS)

10.1.85.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	48.03%	304	TIMSS	CR	0.39	0.42	1.09	0.05	0.000

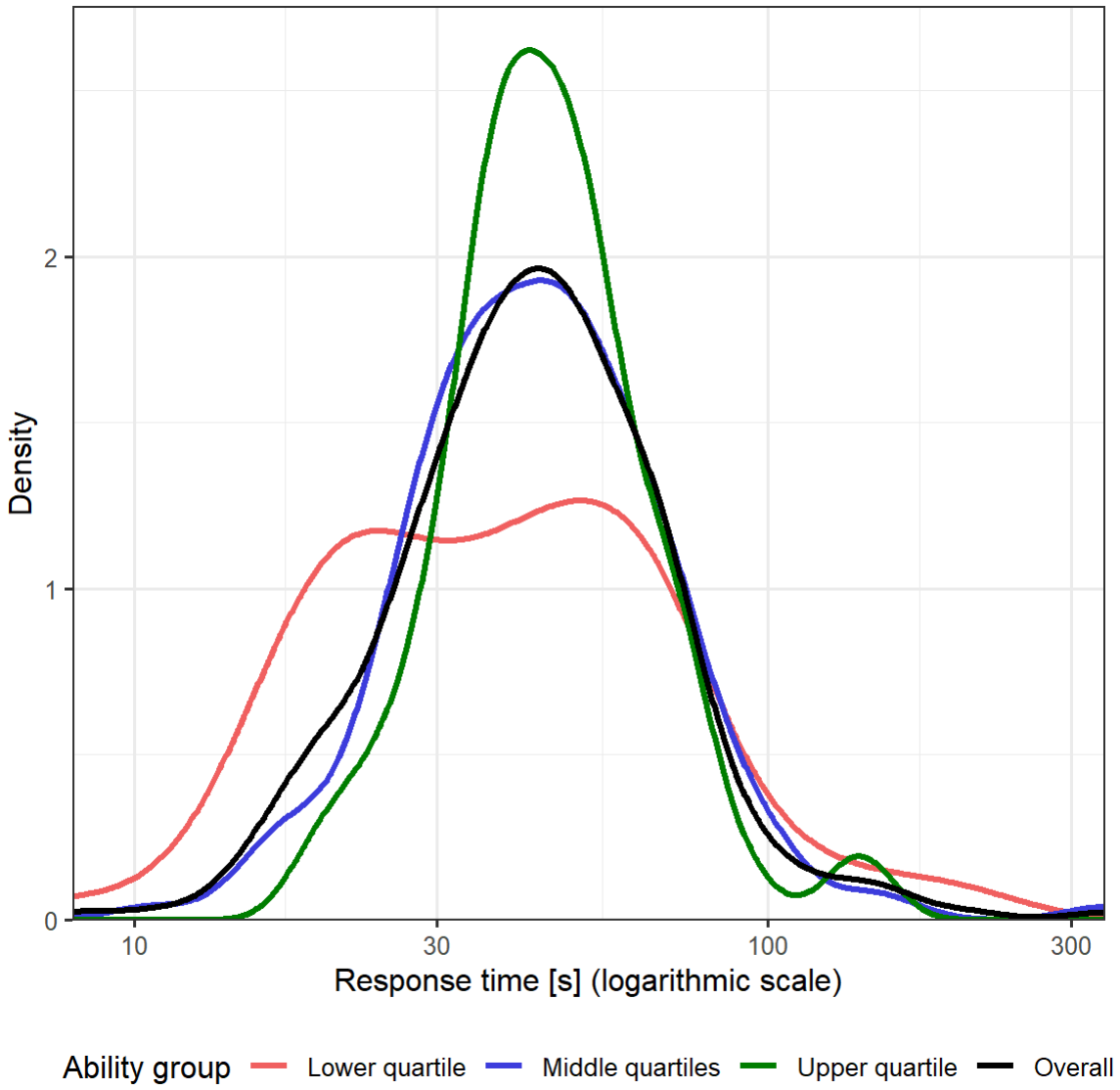
10.1.85.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
10.02	0.007	0.614	0.026	N/A

10.1.85.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

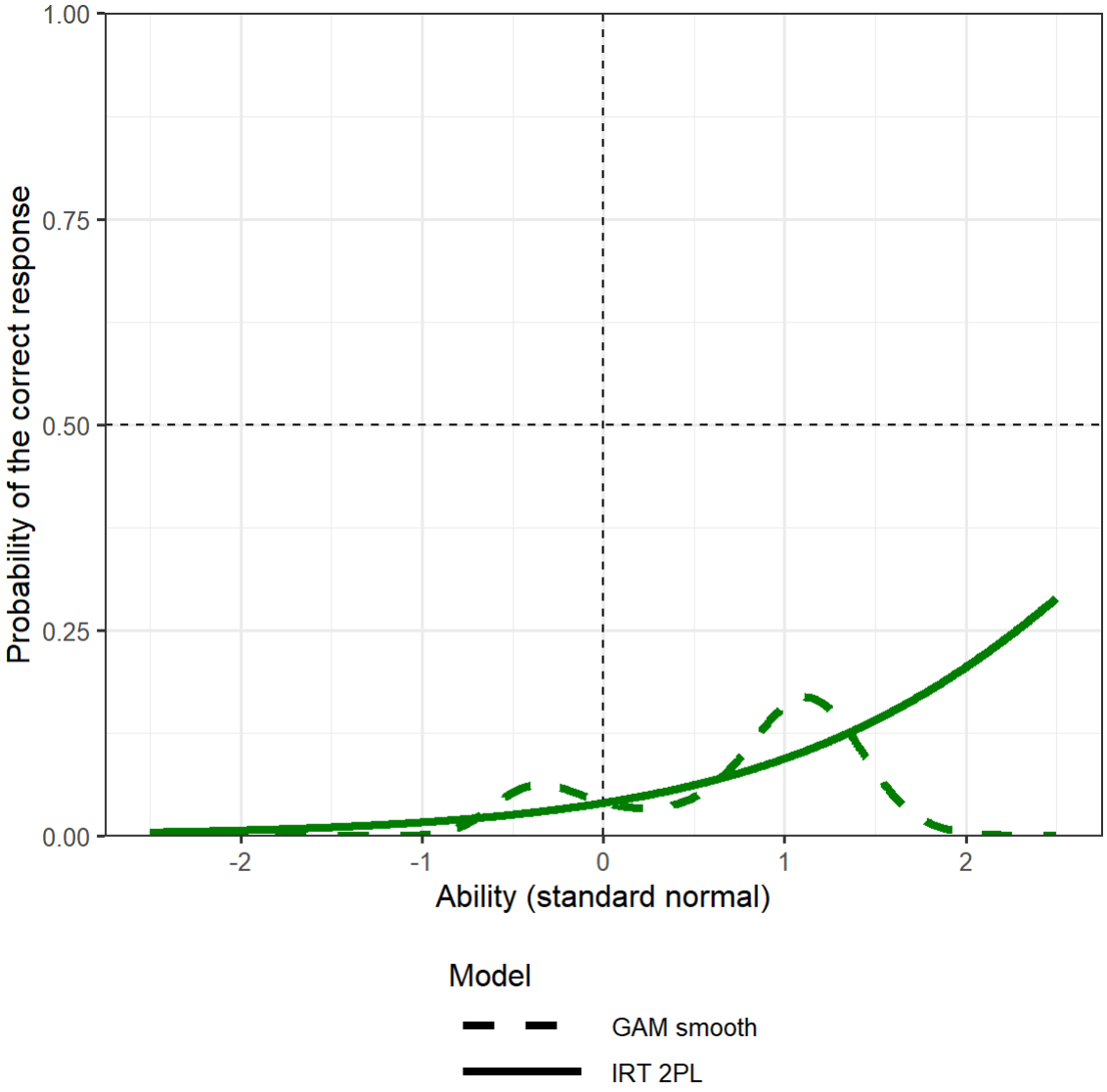
10.1.85.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	8.0	21.6	39.6	62.2	209.0	3.6	0.6
Middle quartiles	10.5	31.7	43.3	57.9	338.6	3.8	0.5
Upper quartile	19.4	35.9	44.6	54.8	143.9	3.8	0.4
Overall	8.0	30.8	43.0	57.4	338.6	3.7	0.5

10.1.86 M890010 (LLM)

10.1.86.1 Item fit:



							IRT model including both TIMSS generated items		
					$r_{score,\theta}$				
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	5.36%	281	LLM complex setup	CR	0.15	0.15	0.91	3.49	0.00

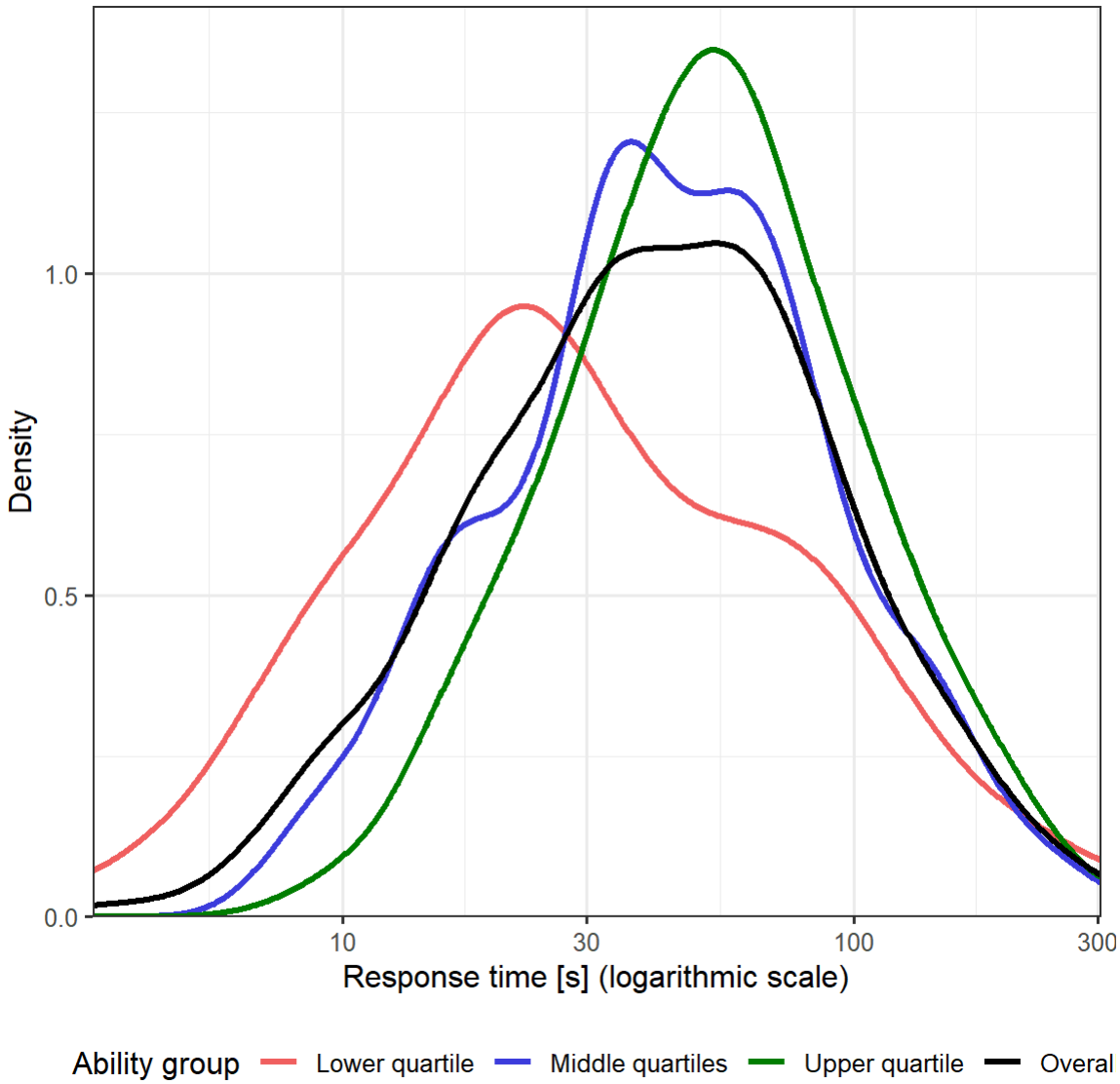
10.1.86.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.80	0.149	1.000	0.034	N/A

10.1.86.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

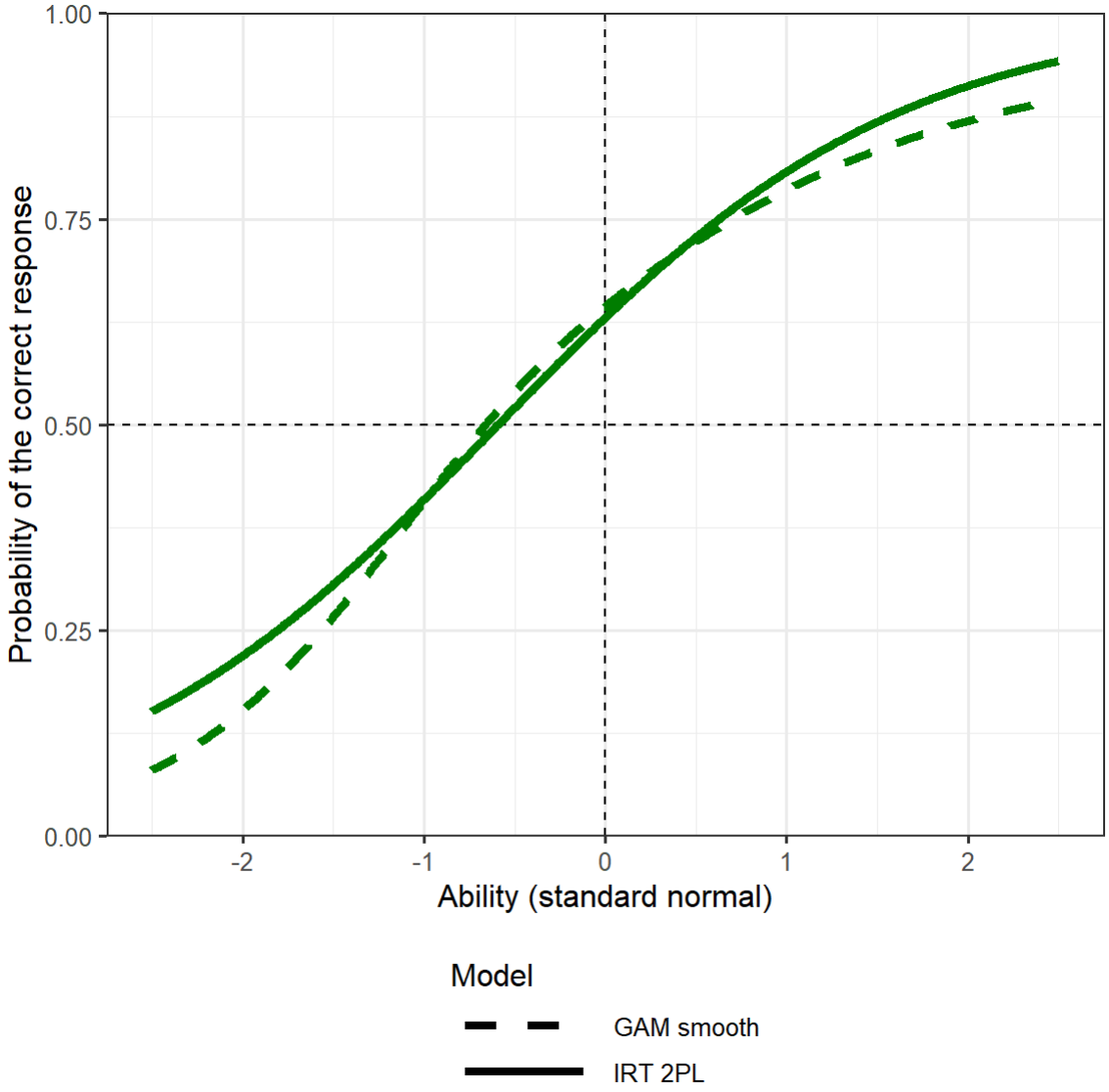
10.1.86.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.2	16.9	26.8	60.5	302.8	3.4	1.0
Middle quartiles	8.3	29.0	40.7	70.7	274.7	3.8	0.7
Upper quartile	9.7	32.2	52.9	79.9	260.0	3.9	0.7
Overall	3.2	22.9	40.6	70.9	302.8	3.7	0.8

10.1.87 M061167 (TIMSS)

10.1.87.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	60.53%	304	TIMSS	CR	0.38	0.35	0.90	−0.59	0.000

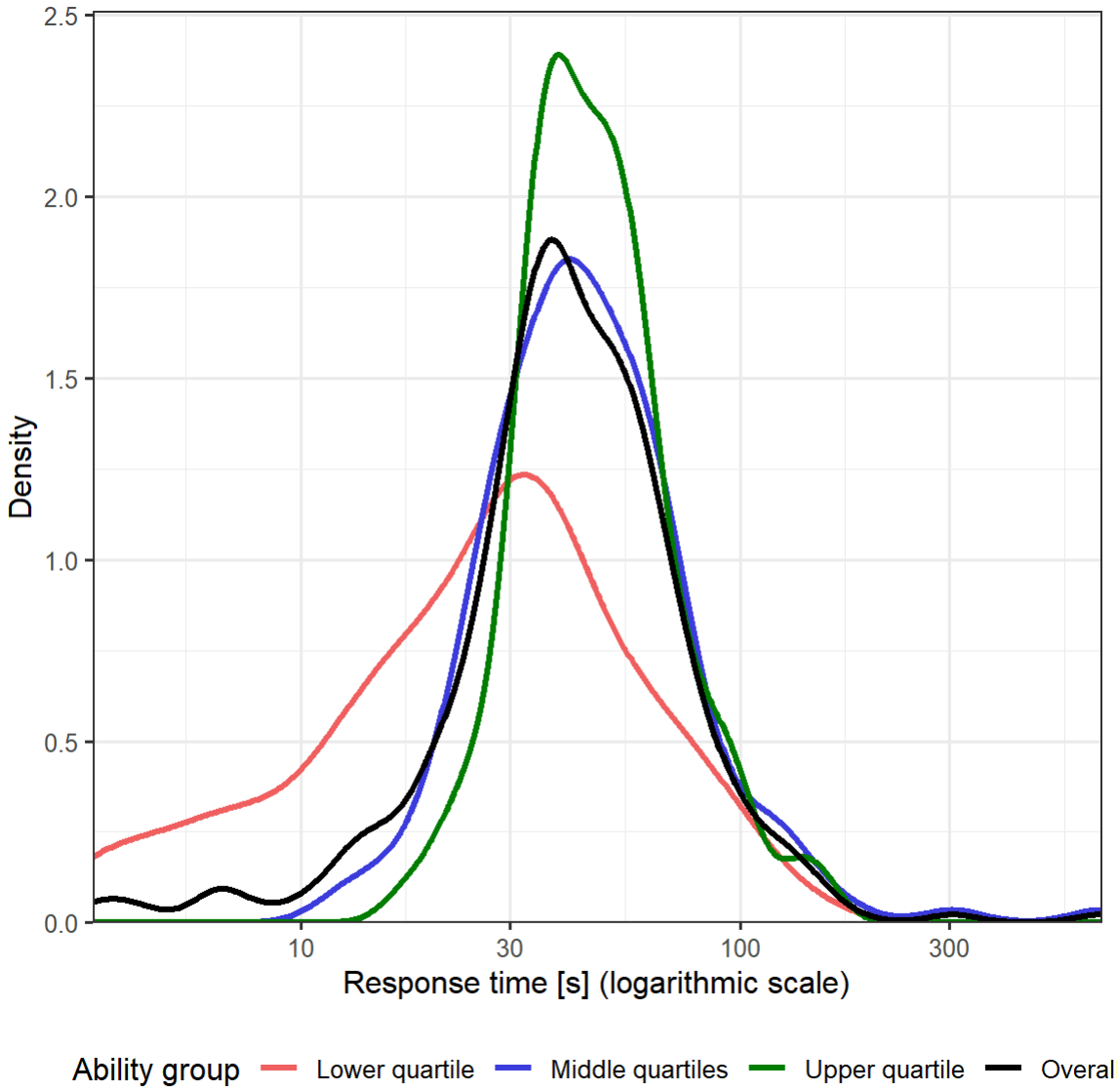
10.1.87.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
7.06	0.029	1.000	0.019	N/A

10.1.87.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

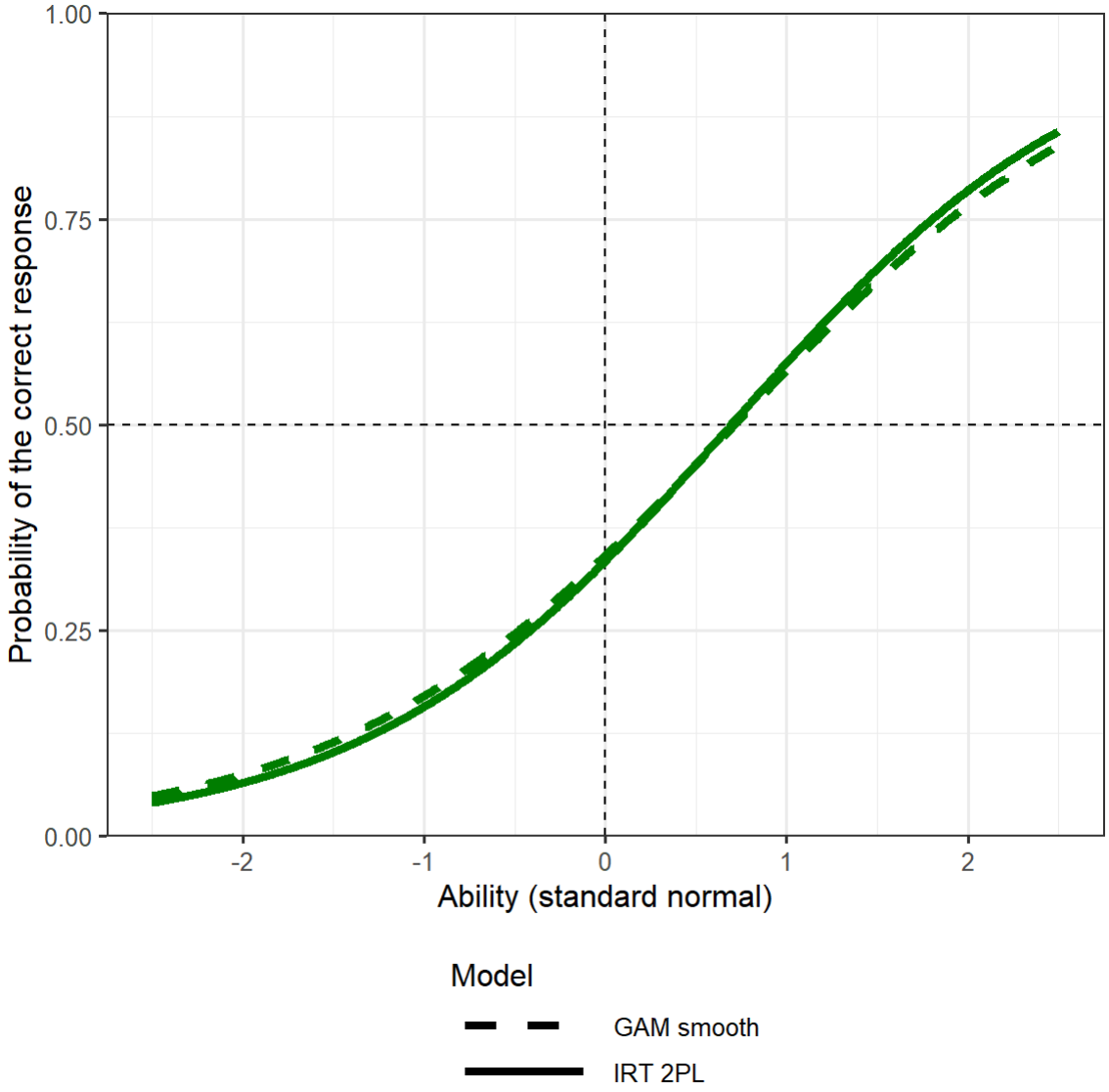
10.1.87.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.4	15.0	30.1	40.6	126.3	3.2	0.8
Middle quartiles	11.9	31.0	41.8	61.1	665.0	3.8	0.6
Upper quartile	17.7	35.9	45.4	58.0	148.1	3.8	0.4
Overall	3.4	30.1	39.6	57.3	665.0	3.7	0.6

10.1.88 M041276 (TIMSS)

10.1.88.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	37.07%	294	TIMSS	CR	0.38	0.46	0.99	0.70	0.000

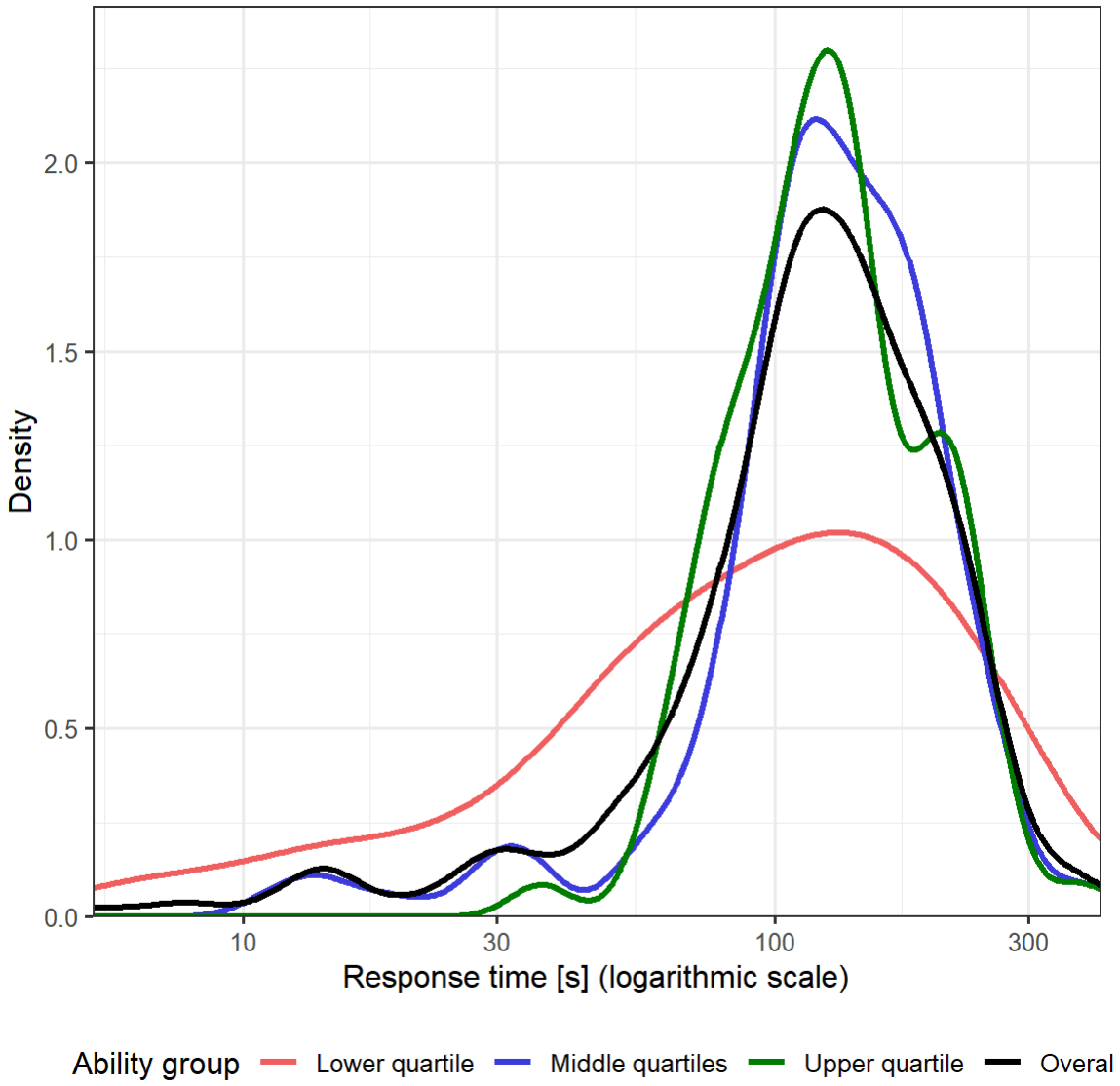
10.1.88.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.06	0.972	1.000	0.000	N/A

10.1.88.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

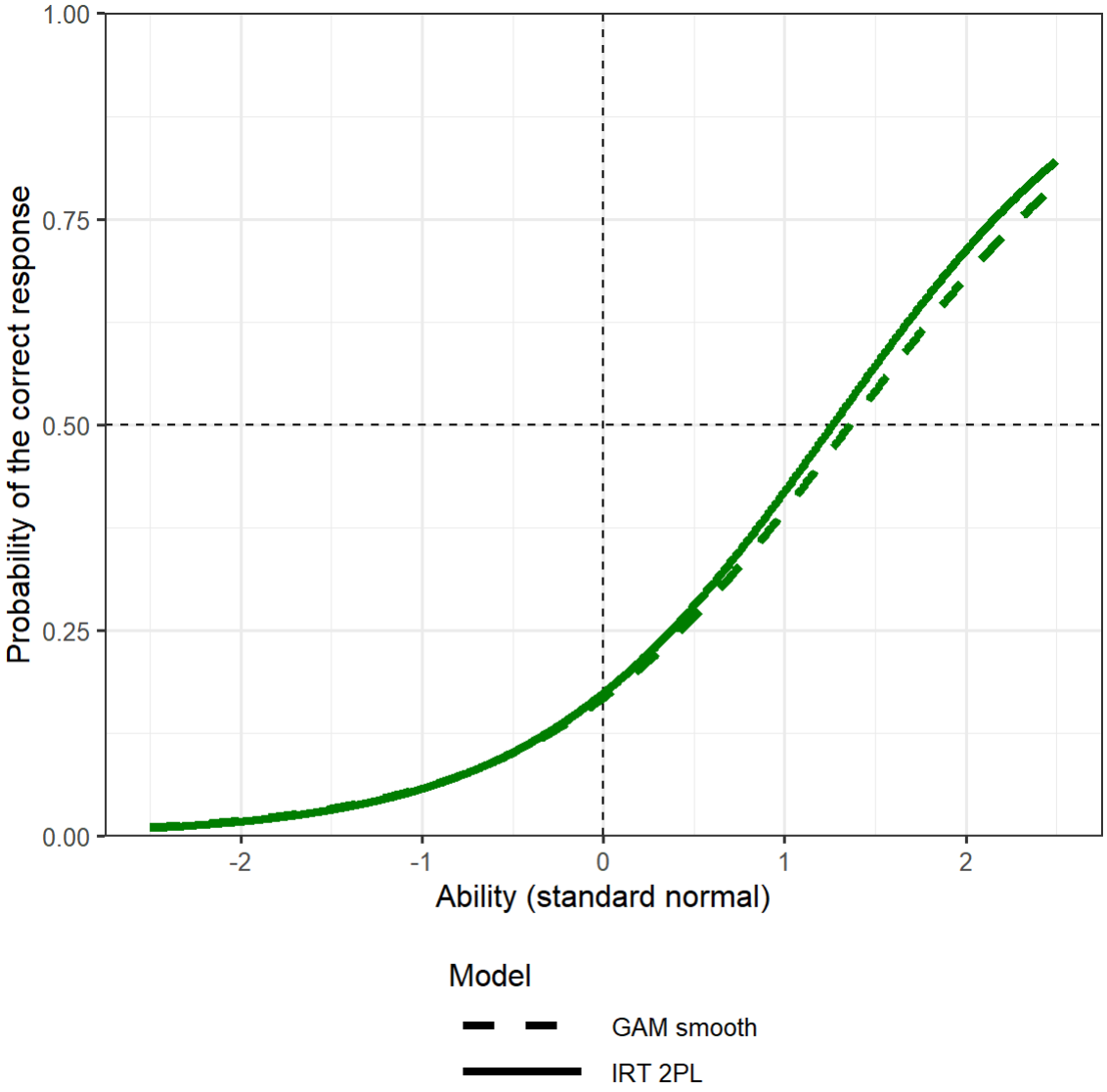
10.1.88.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.2	49.8	100.6	164.7	368.4	4.4	0.9
Middle quartiles	11.4	100.7	129.1	179.0	410.0	4.8	0.6
Upper quartile	36.4	97.5	129.2	157.9	380.1	4.8	0.4
Overall	5.2	88.1	122.5	172.2	410.0	4.7	0.7

10.1.89 M061236 (TIMSS)

10.1.89.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	22.04%	304	TIMSS	CR	0.40	0.41	1.23	1.27	0.019

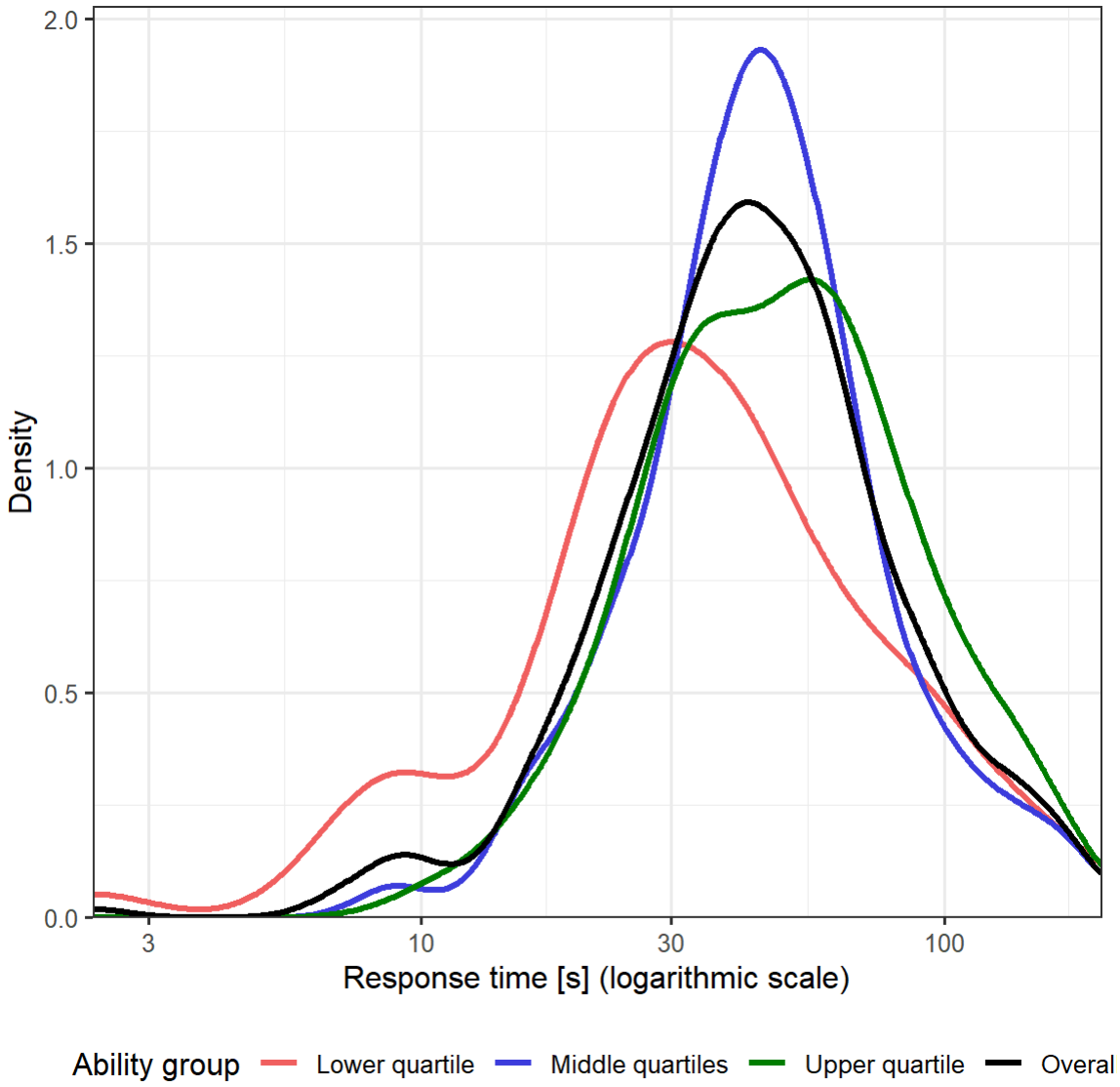
10.1.89.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.51	0.173	1.000	0.012	N/A

10.1.89.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

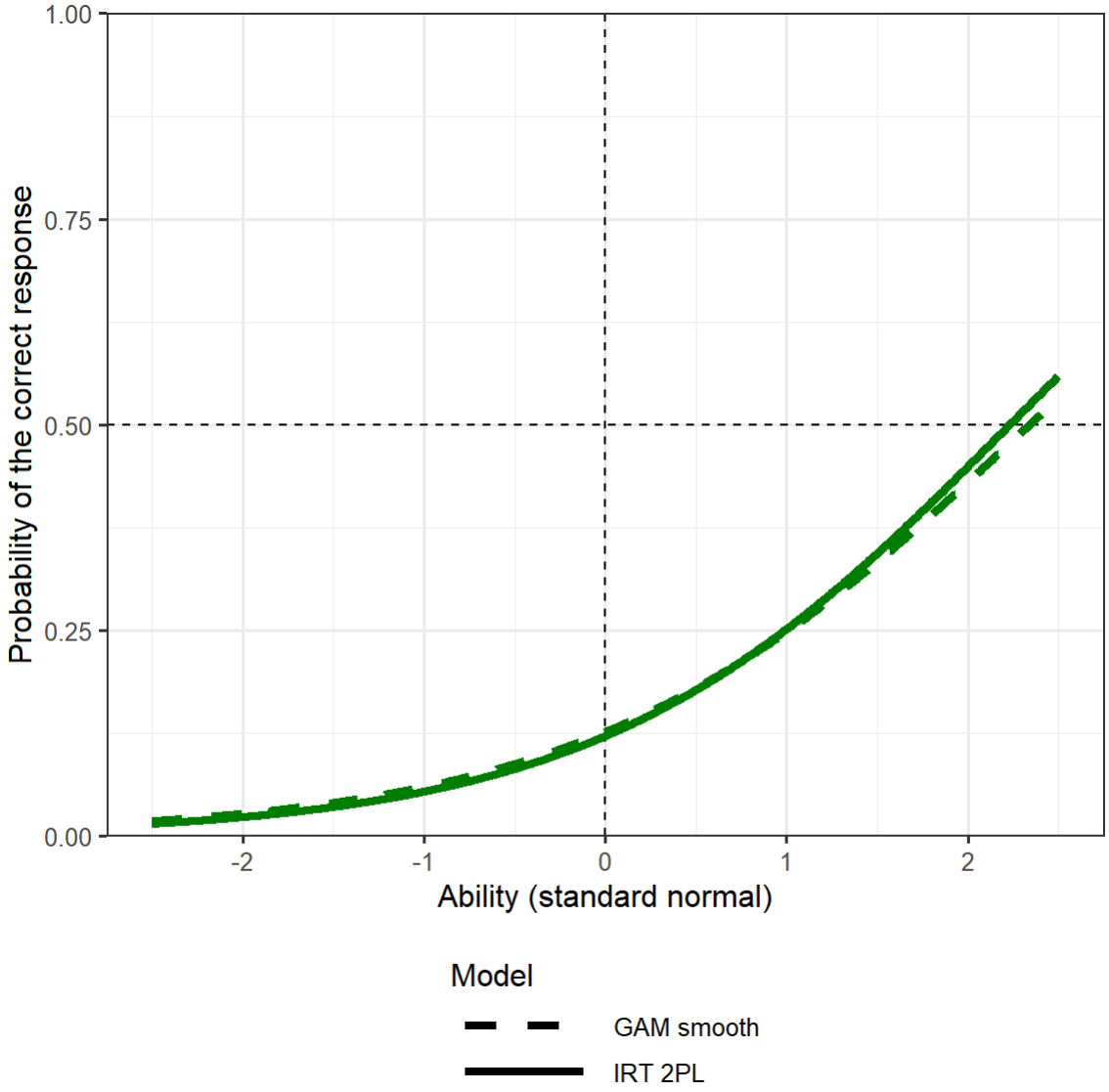
10.1.89.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.4	22.2	35.6	55.1	184.2	3.5	0.8
Middle quartiles	8.3	32.6	44.0	60.9	199.7	3.8	0.6
Upper quartile	10.6	33.0	50.1	73.5	167.1	3.9	0.6
Overall	2.4	29.2	42.8	62.2	199.7	3.8	0.6

10.1.90 M051236 (TIMSS)

10.1.90.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	15.65%	294	TIMSS	CR	0.27	0.27	0.89	2.23	0.000

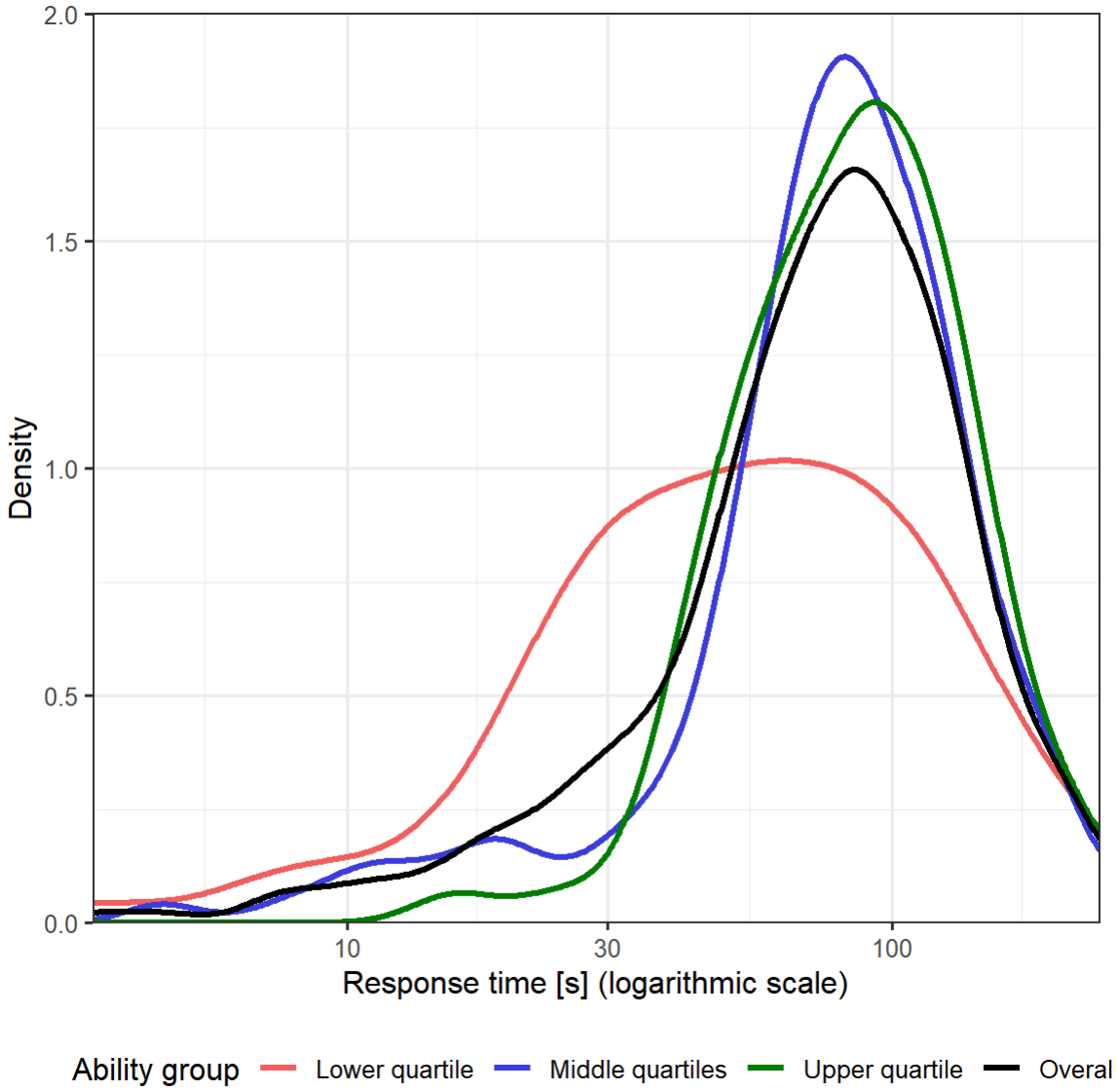
10.1.90.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.56	0.278	1.000	0.011	N/A

10.1.90.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

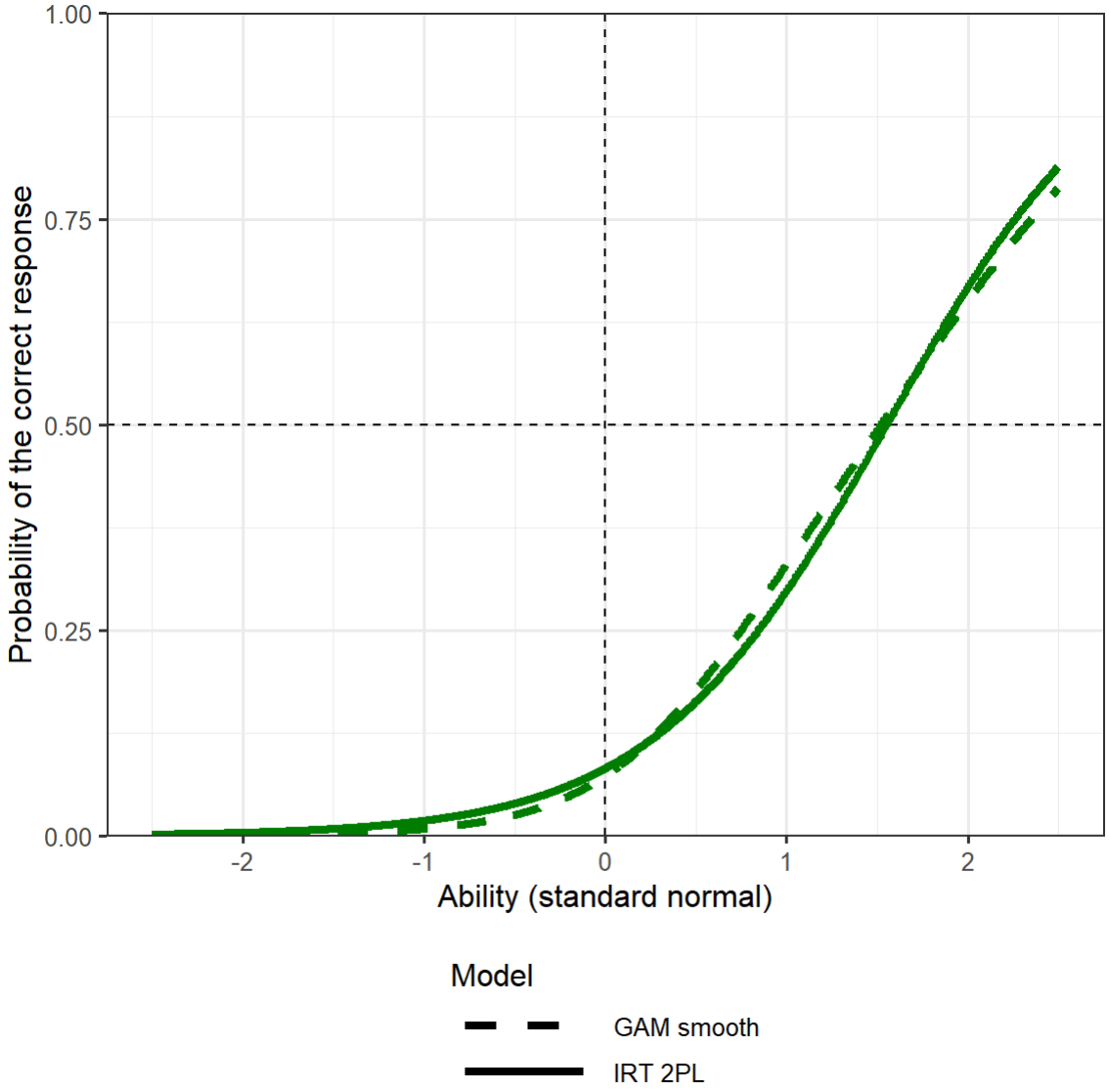
10.1.90.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.4	31.0	54.0	92.0	230.4	3.9	0.8
Middle quartiles	4.6	59.0	81.6	113.7	240.7	4.3	0.7
Upper quartile	15.8	59.4	86.2	116.8	231.2	4.4	0.5
Overall	3.4	52.3	77.9	110.2	240.7	4.2	0.7

10.1.91 M100027 (LLM)

10.1.91.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS a generated items Discrimination	Item difficulty	RMSEA
1	15.67%	319	LLM simple setup	CR	0.47	0.25	1.55	1.55	0.000

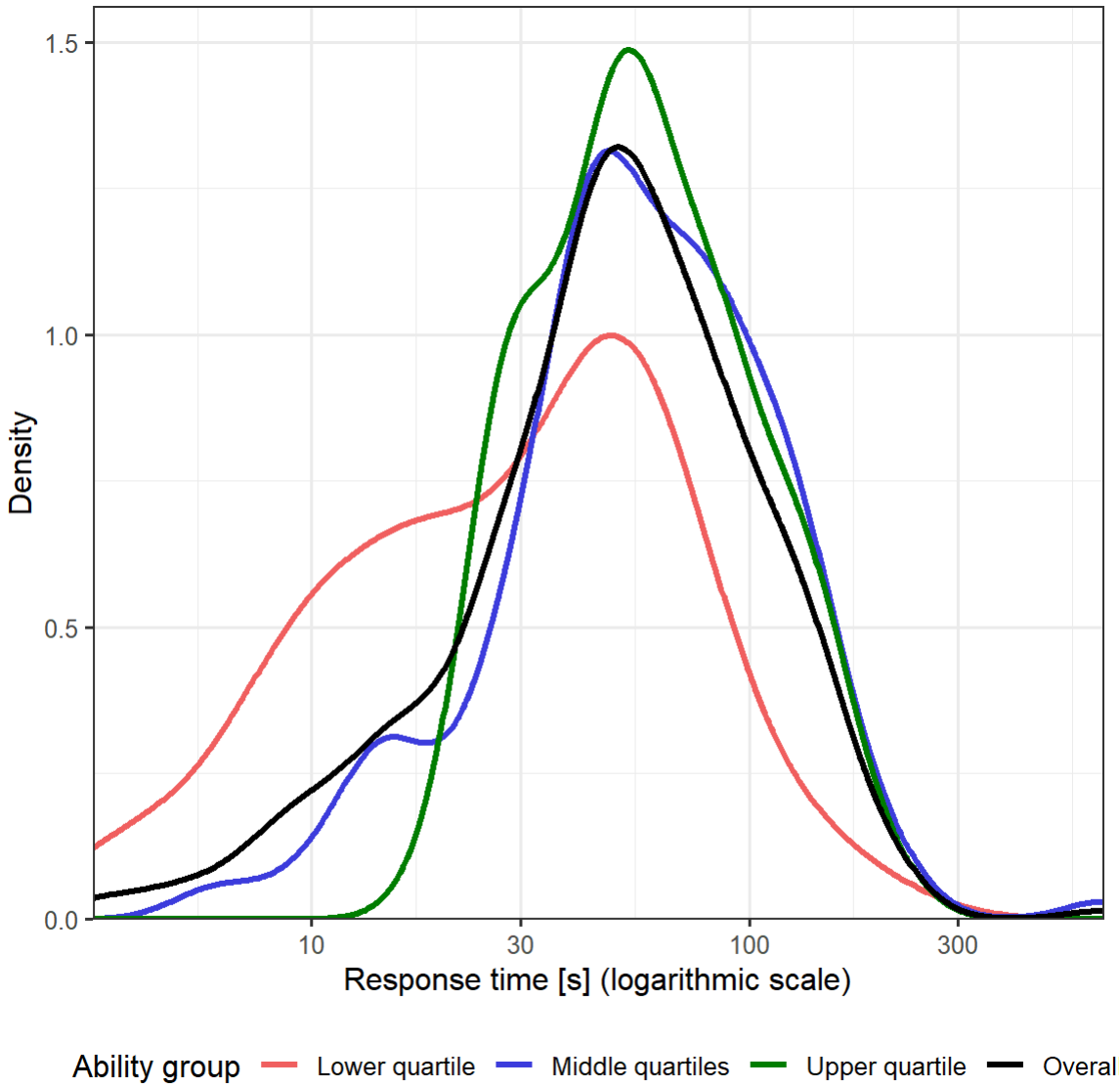
10.1.91.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.49	0.783	1.000	0.002	N/A

10.1.91.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

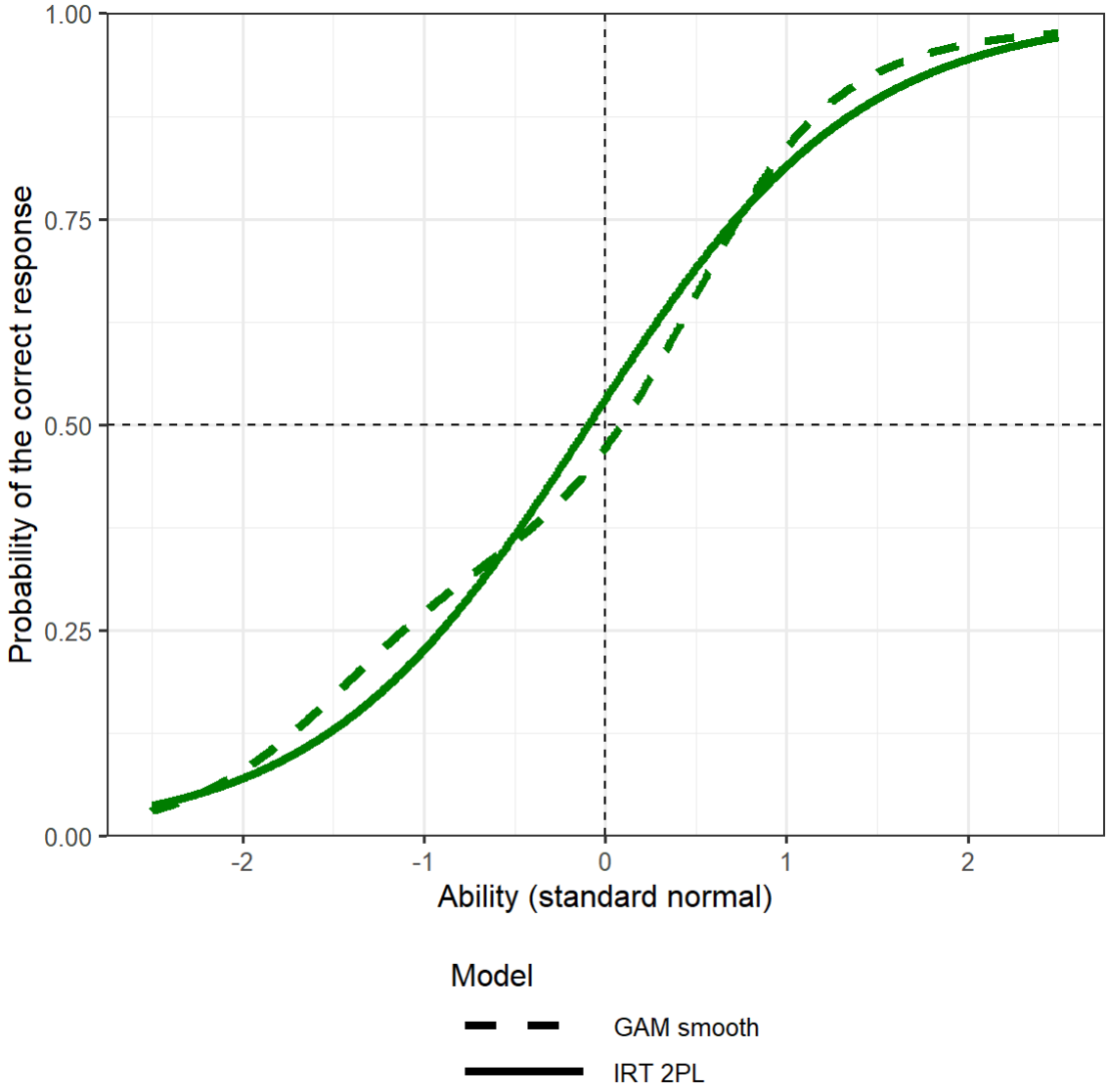
10.1.91.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.2	14.2	34.0	53.9	206.1	3.3	0.9
Middle quartiles	5.6	38.1	56.2	88.3	645.0	4.0	0.7
Upper quartile	18.9	36.9	55.8	85.5	215.3	4.1	0.6
Overall	3.2	29.8	49.8	79.0	645.0	3.8	0.8

10.1.92 MP71217 (TIMSS)

10.1.92.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	51.15%	306	TIMSS	CR	0.50	0.51	1.35	−0.09	0.007

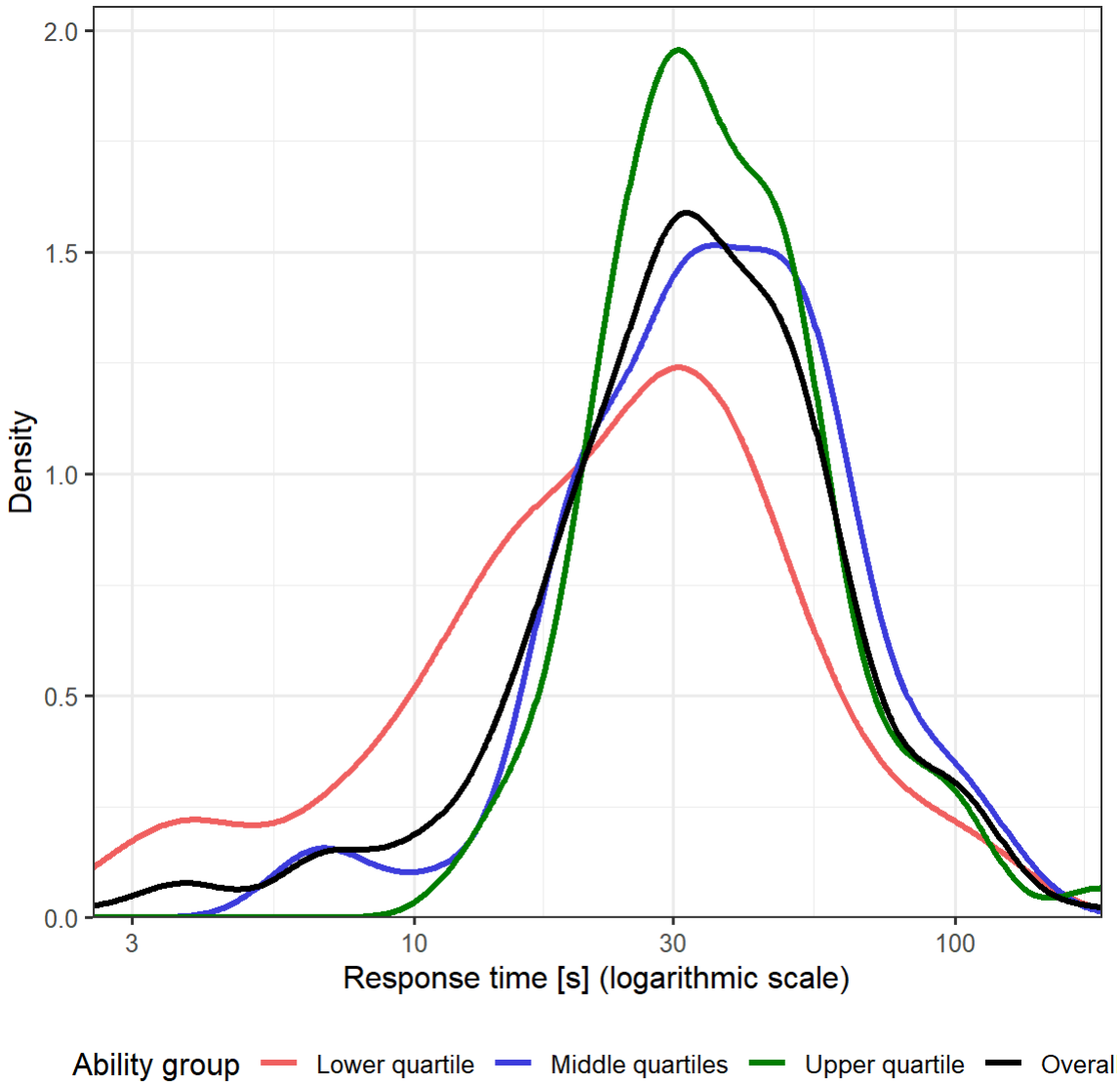
10.1.92.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
8.56	0.014	1.000	0.022	N/A

10.1.92.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

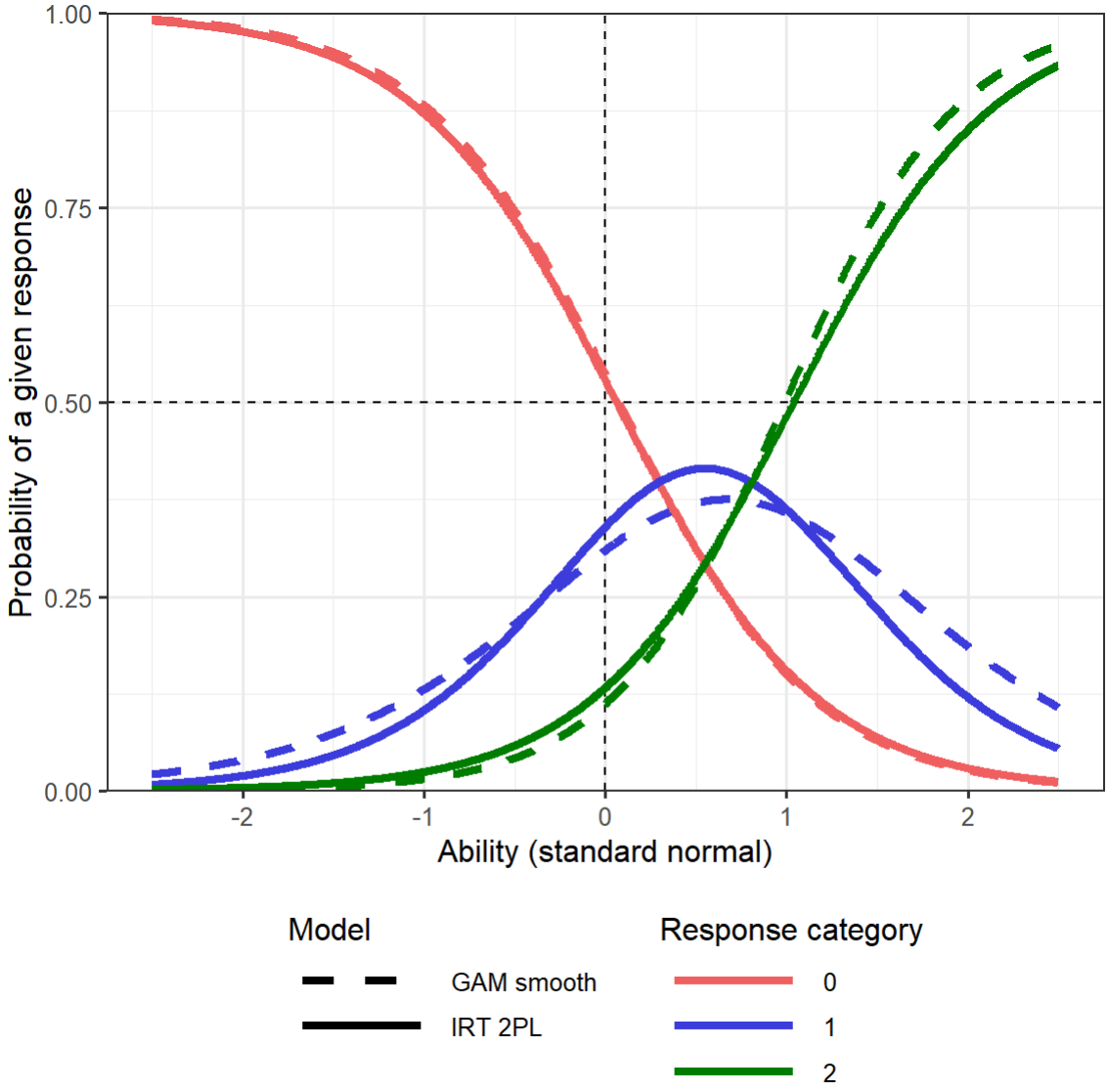
10.1.92.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.5	14.5	26.0	38.5	119.7	3.1	0.8
Middle quartiles	5.8	24.3	35.2	51.1	138.7	3.6	0.6
Upper quartile	12.5	25.9	34.5	47.7	186.2	3.6	0.5
Overall	2.5	21.8	32.3	48.4	186.2	3.4	0.7

10.1.93 M061248 (TIMSS)

10.1.93.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	Catego difficul
2	34.26%	306	TIMSS	CR	0.64	0.62	1.81	0.55	0.0

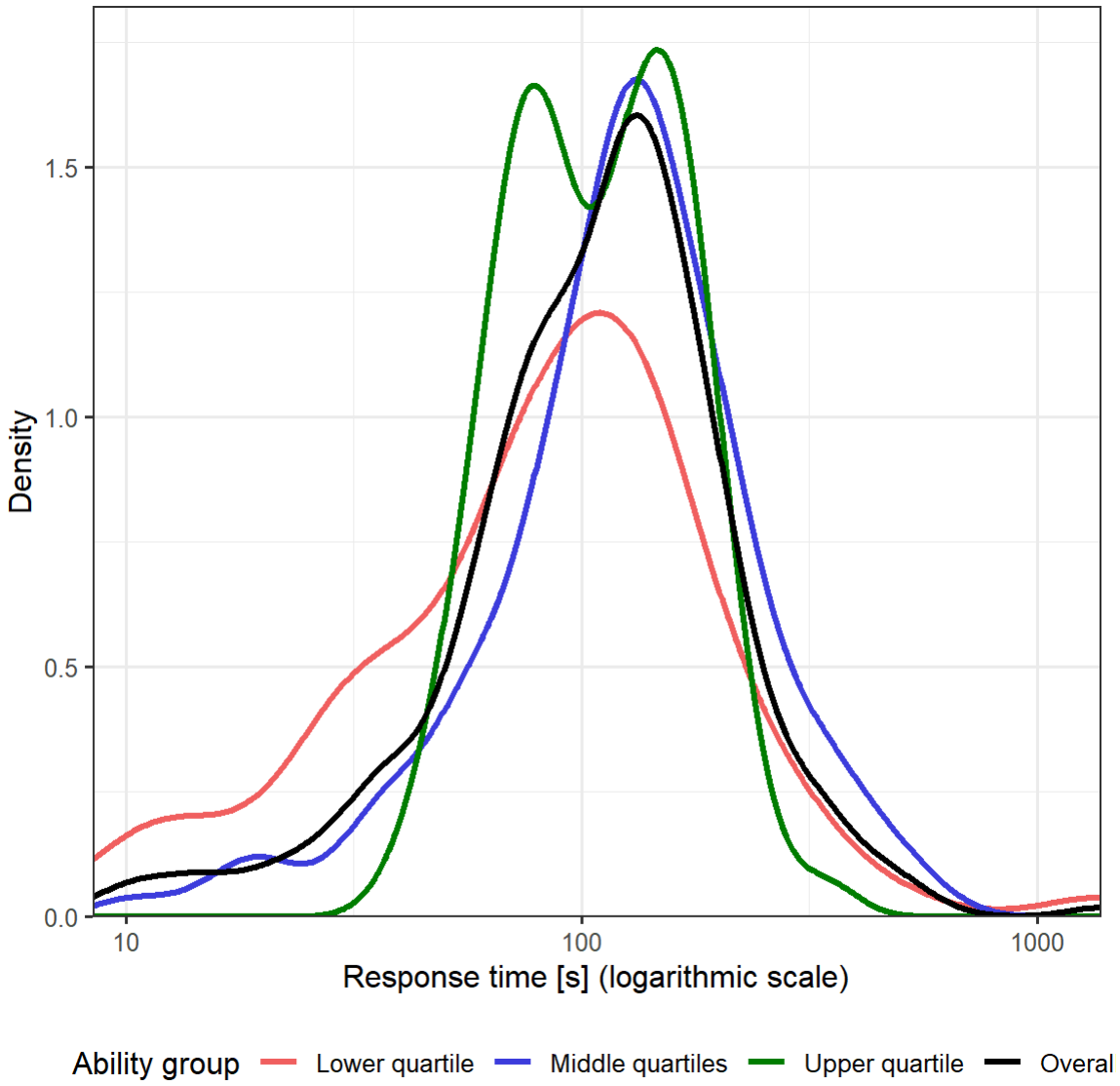
10.1.93.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
11.59	0.003	0.286	0.028	N/A

10.1.93.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

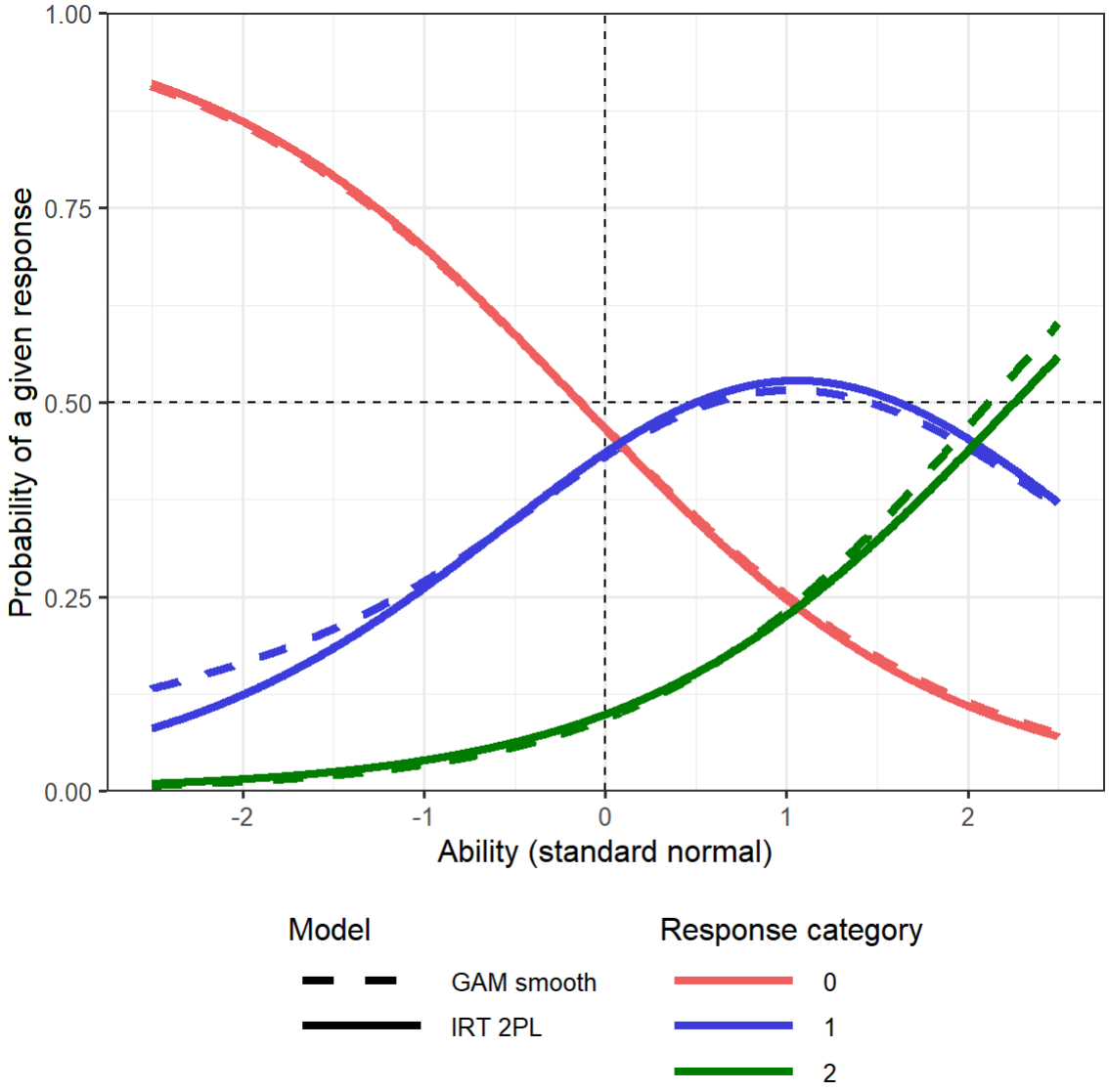
10.1.93.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	8.4	48.3	91.5	137.3	1,375.1	4.4	0.9
Middle quartiles	10.2	87.1	127.0	188.0	588.6	4.8	0.7
Upper quartile	40.6	74.2	113.4	154.4	342.1	4.7	0.4
Overall	8.4	72.4	116.1	159.3	1,375.1	4.6	0.7

10.1.94 M890009 (LLM)

10.1.94.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIM Discrimination	IRT model including both TIM Item difficulty	IRT model including both TIM Categ difficu
2	33.23%	319	LLM complex setup	CR	0.45	0.26	0.98	1.06	−0

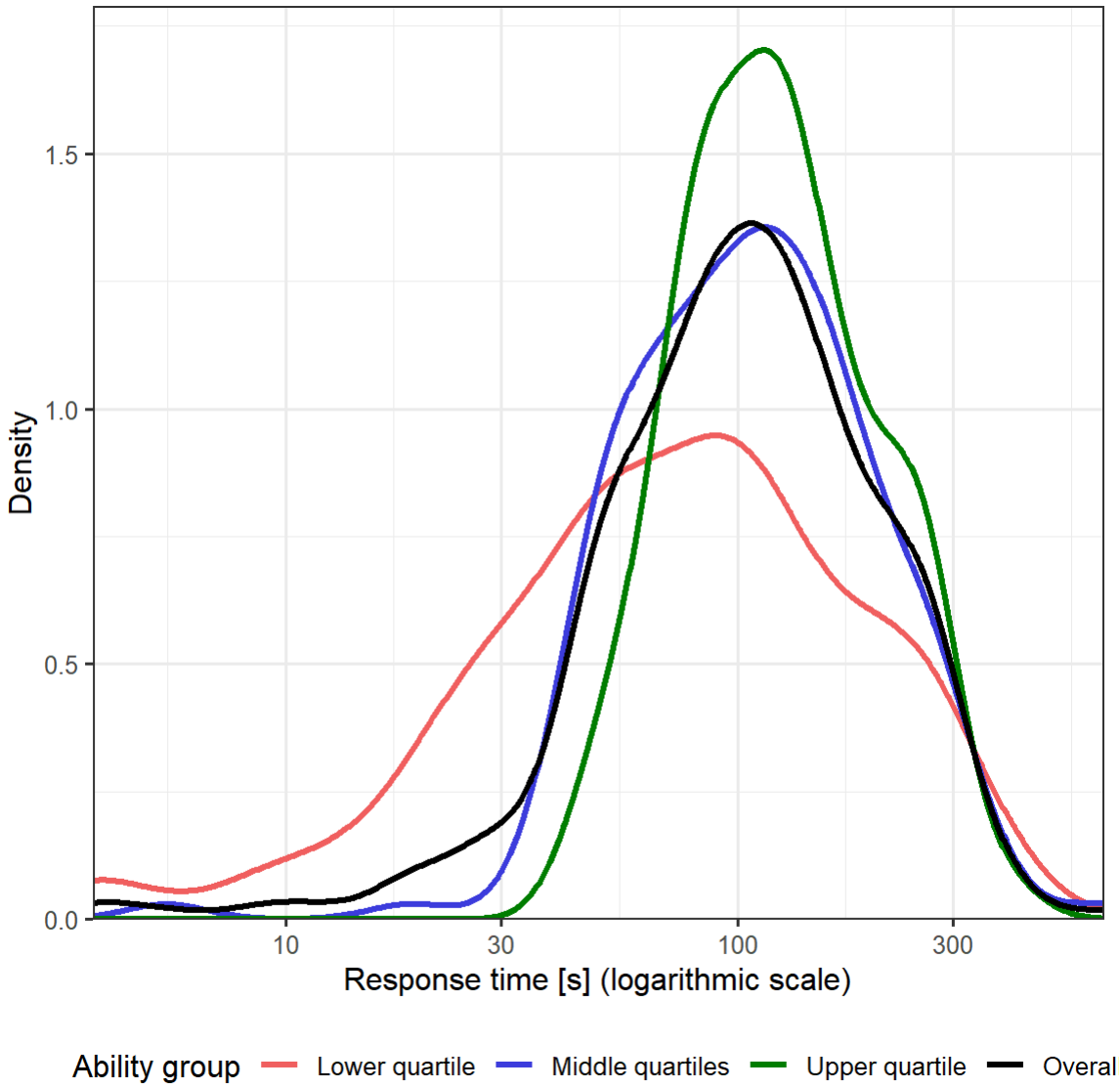
10.1.94.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.78	0.678	1.000	0.002	N/A

10.1.94.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

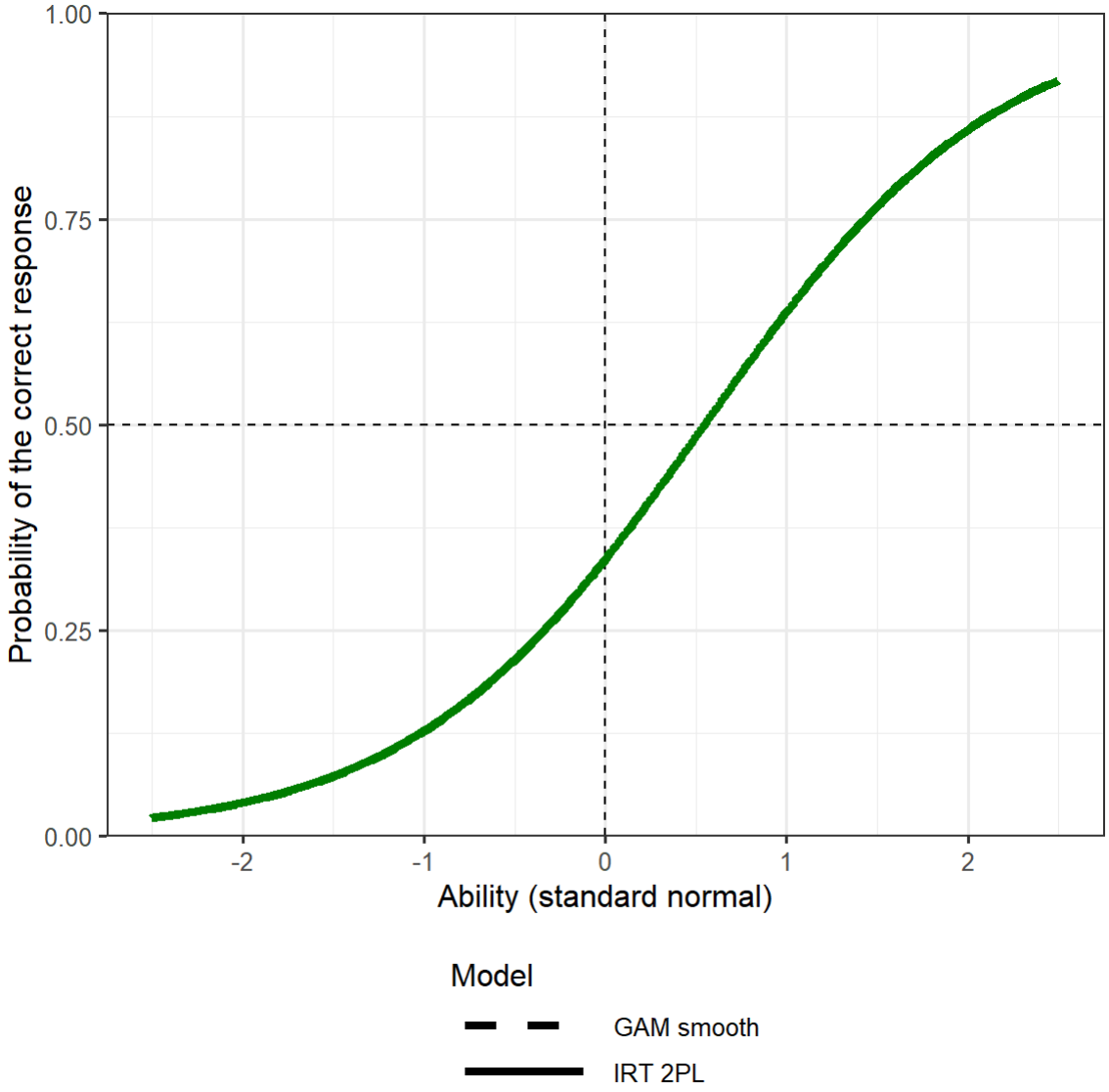
10.1.94.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.8	43.4	76.3	125.9	409.3	4.2	1.0
Middle quartiles	5.5	66.6	108.2	161.1	644.7	4.7	0.6
Upper quartile	46.4	83.8	123.0	167.0	399.6	4.8	0.5
Overall	3.8	65.1	104.0	159.2	644.7	4.6	0.7

10.1.95 M890005 (LLM)

10.1.95.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	36.79%	281	LLM complex setup	CR	0.45	0.32	1.24	0.55	0.02

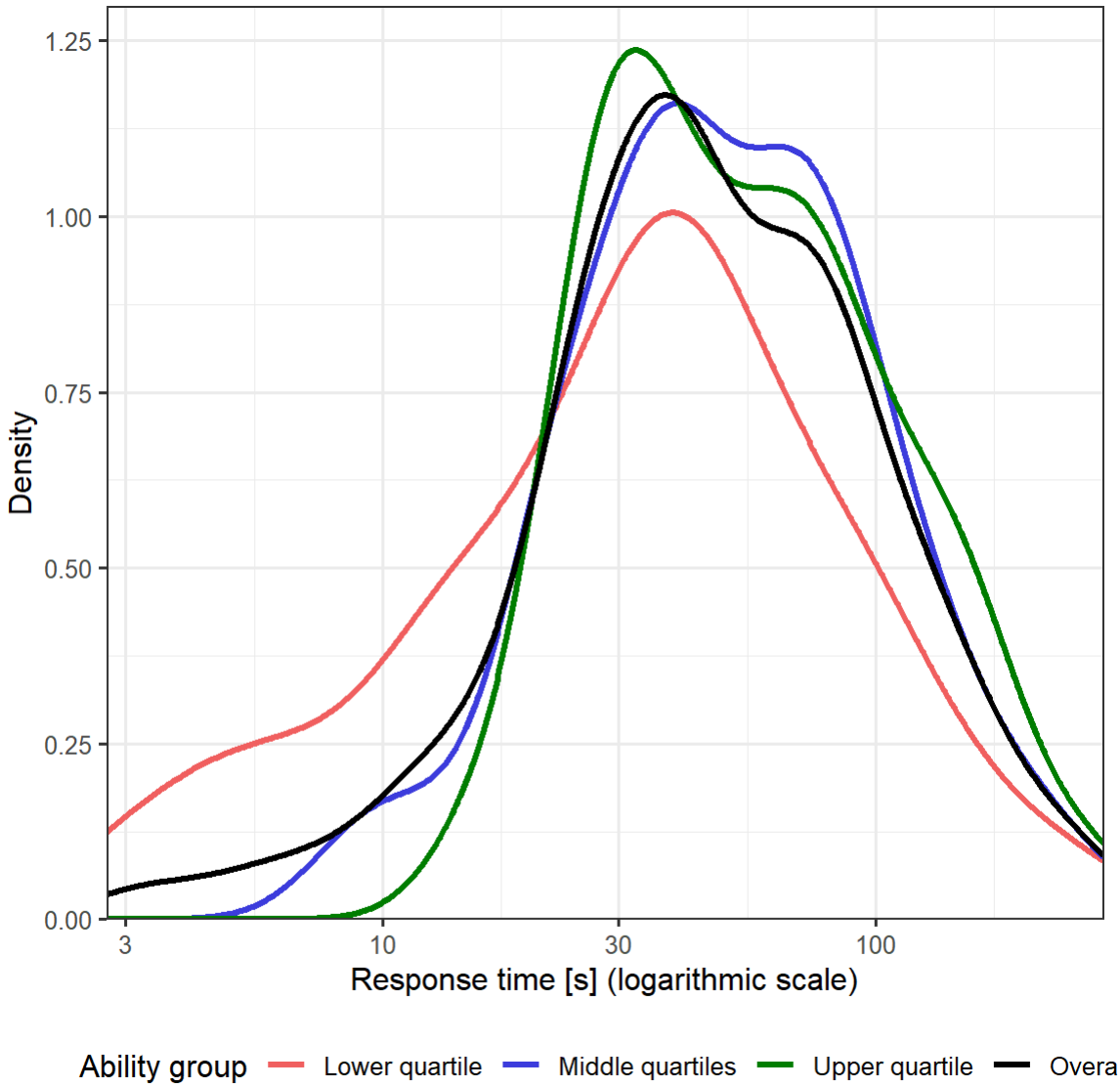
10.1.95.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
11.32	0.003	0.324	0.033	N/A

10.1.95.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

10.1.95.4 Answering time:

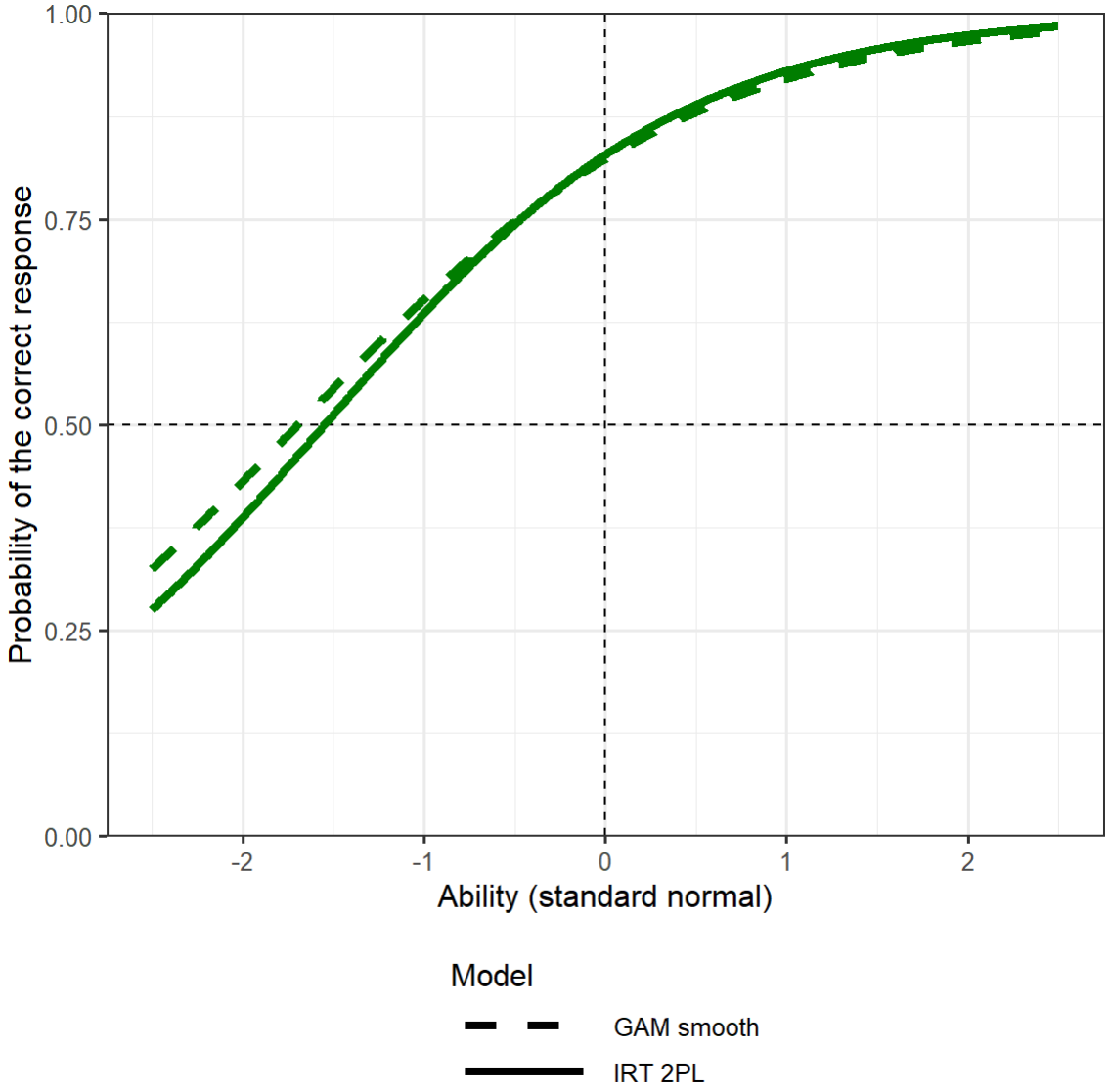


	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.8	17.3	36.0	59.0	264.7	3.4	1.0
Middle quartiles	7.8	30.3	46.7	80.4	278.2	3.9	0.7
Upper quartile	15.4	31.4	54.0	85.3	288.9	4.0	0.7
Overall	2.8	27.3	43.6	78.4	288.9	3.8	0.8

10.2 Science items

10.2.1 S090041 (LLM)

10.2.1.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	79.19%	299	LLM teachers' selection	MC	0.32	0.21	1.02	−1.55	0.0

10.2.1.2 Gender DIF:

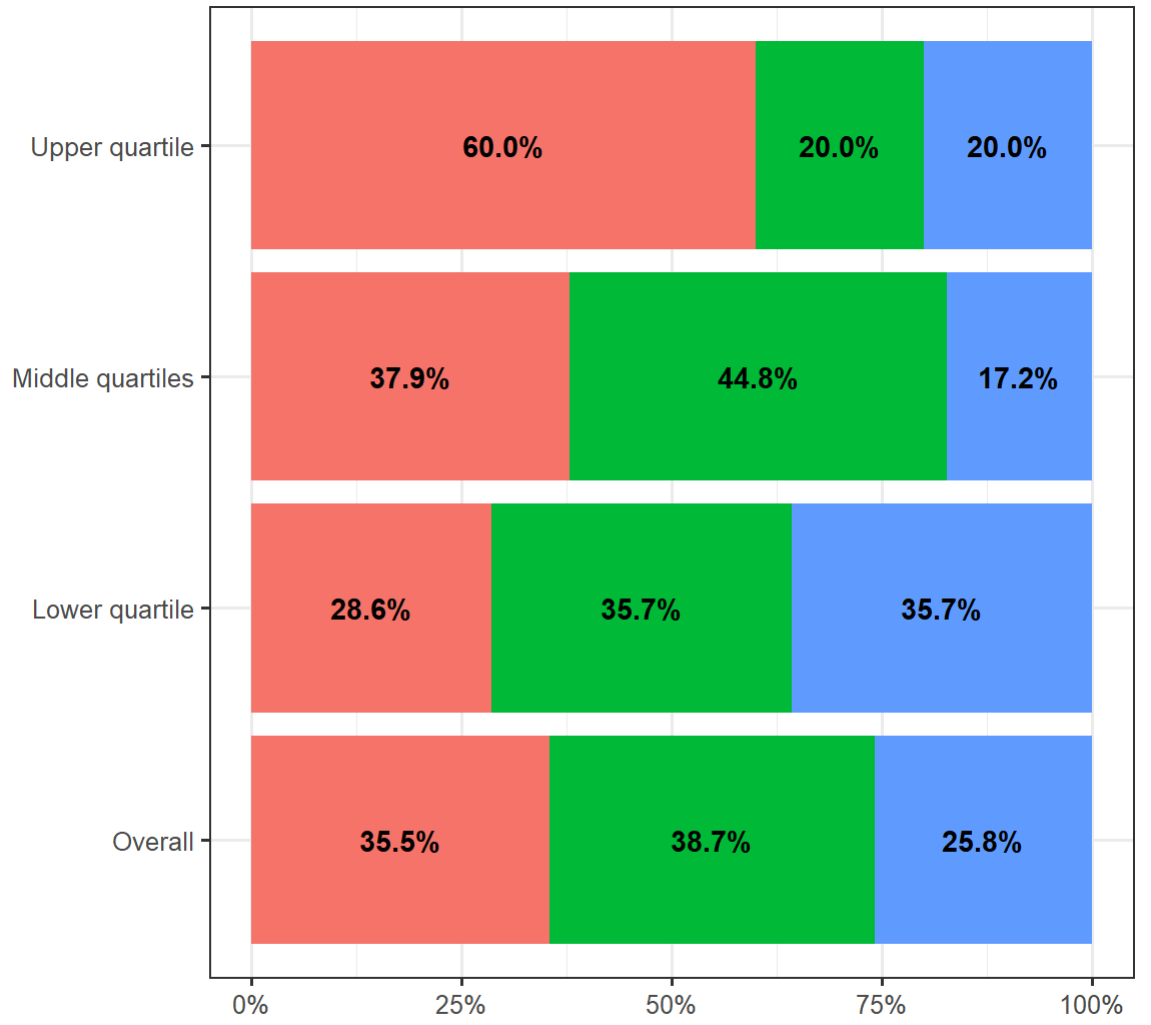
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
6.67	0.036	1.000	0.024	N/A

10.2.1.3 Frequency of answers:

Answer	Frequency	Percent
Cukier rozpuści się w tym samym tempie w obu kubkach.	22	7.4%
Cukier w ciepłej wodzie rozpuści się szybciej. (correct)	236	79.2%
Cukier w obu kubkach nie rozpuści się.	24	8.1%
Cukier w zimnej wodzie rozpuści się szybciej.	16	5.4%
Overall	298	100.0%

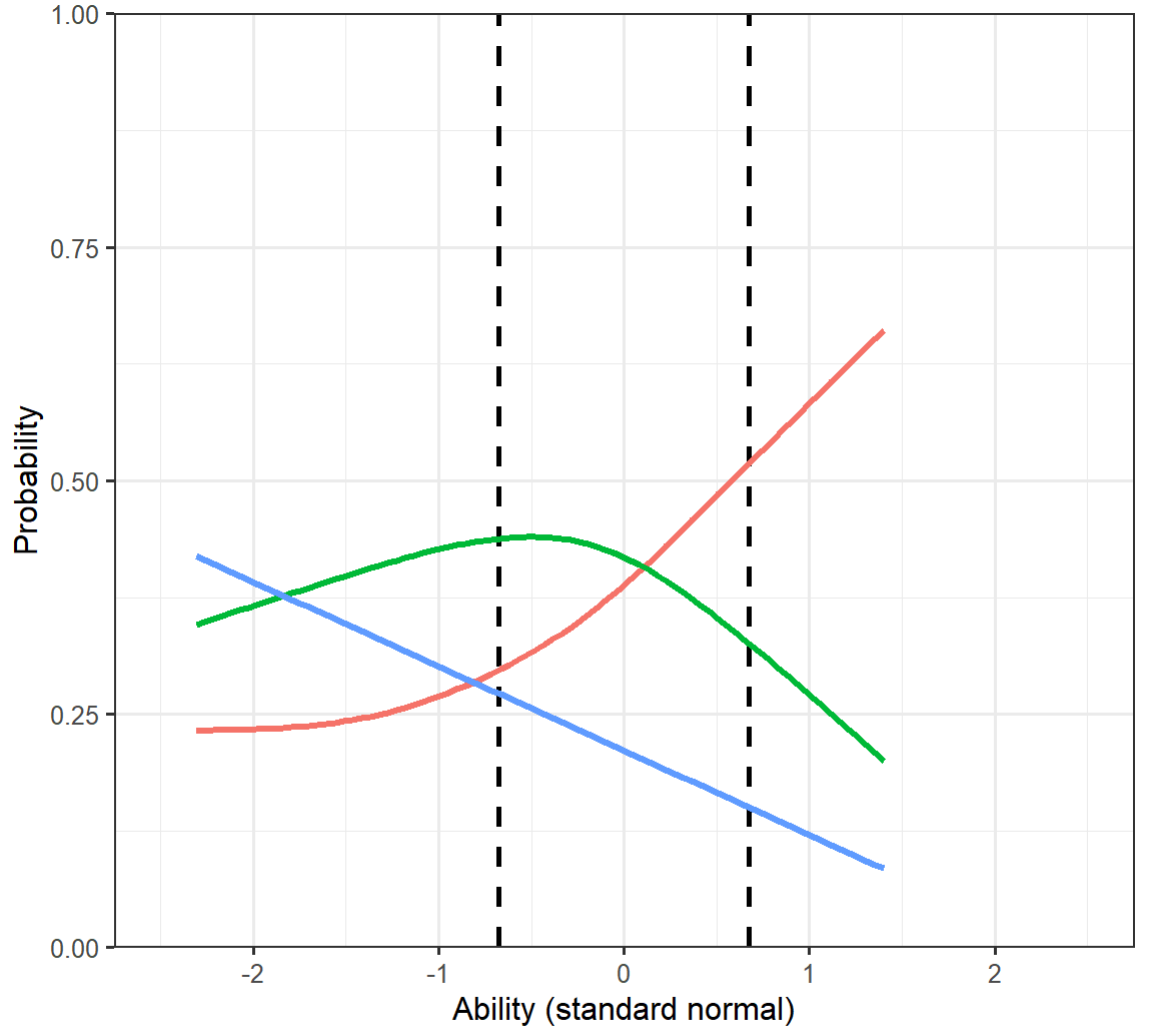
10.2.1.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Cukier rozpuści się w tym samym tempie w obu kubkach.	8	11	3	22	28.6%	37.9%	60.0%	35.5%
Cukier w obu kubkach nie rozpuści się.	10	13	1	24	35.7%	44.8%	20.0%	38.7%
Cukier w zimnej wodzie rozpuści się szybciej.	10	5	1	16	35.7%	17.2%	20.0%	25.8%
Overall	28	29	5	62	100.0%	100.0%	100.0%	100.0%



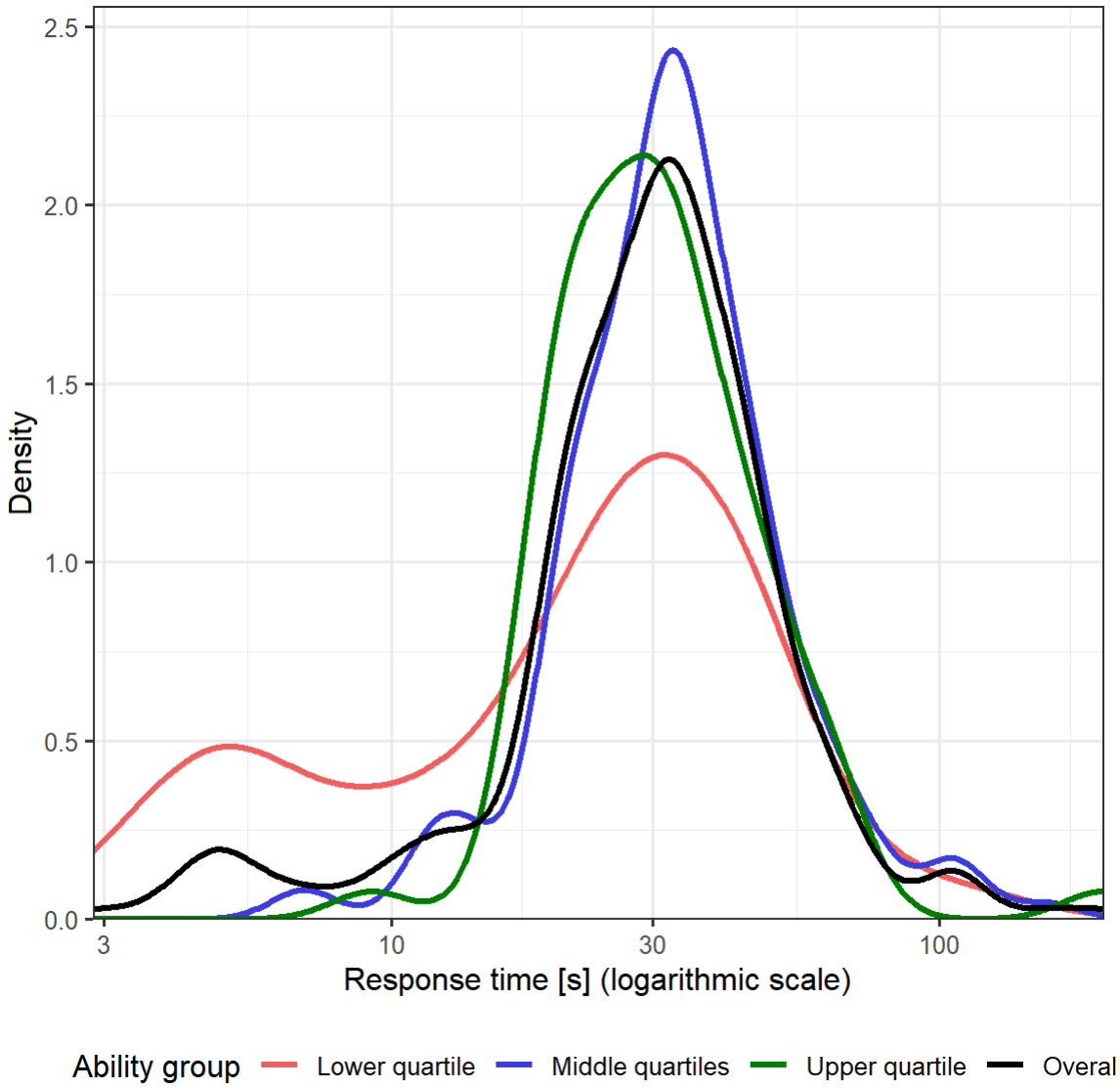
w tym samym tempie w obu kubkach. Cukier w obu kubkach nie rozpuści się. Cukier w z

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



w tym samym tempie w obu kubkach. Cukier w obu kubkach nie rozpuści się. Cukier w z

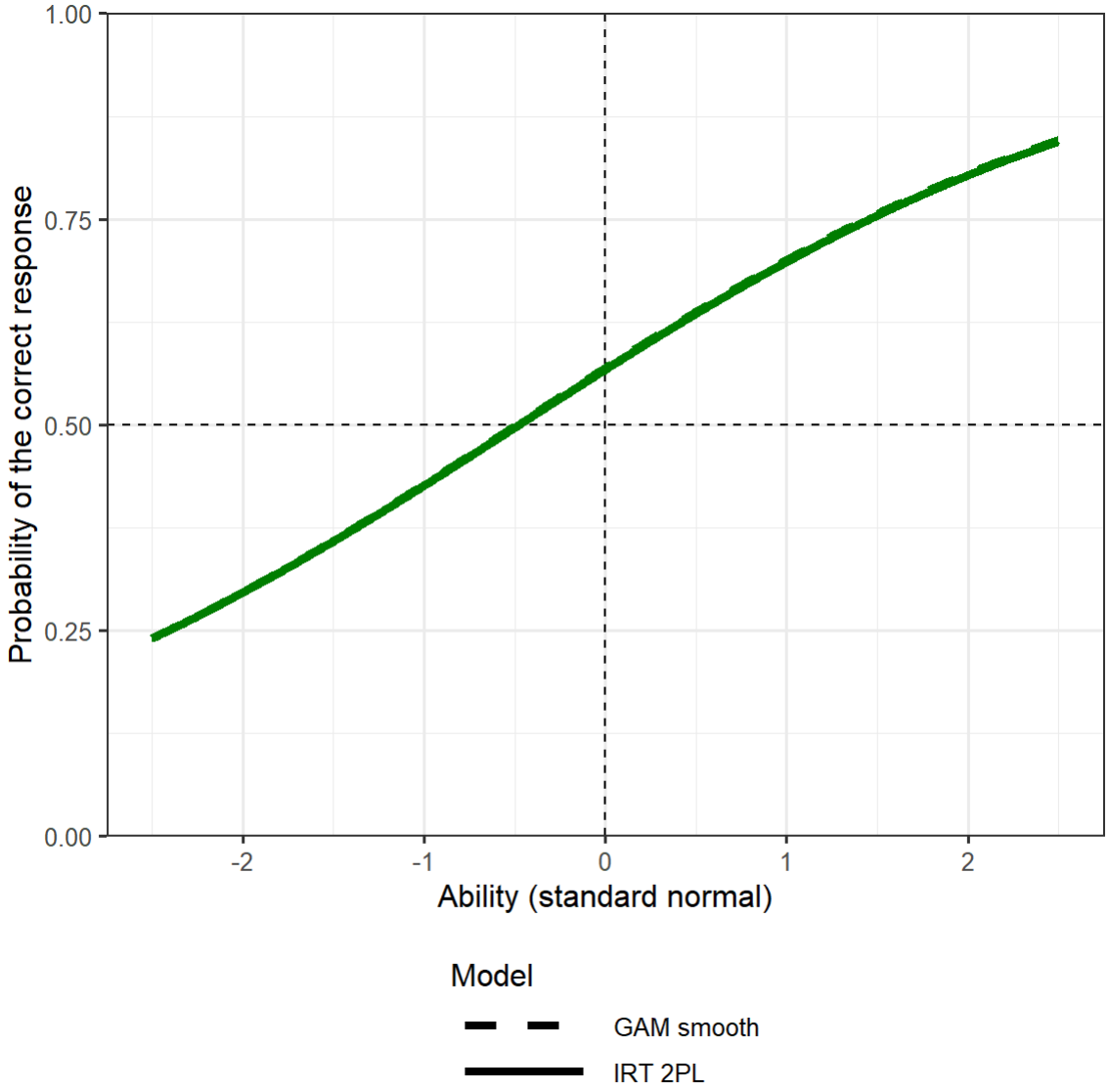
10.2.1.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.9	11.2	25.2	37.6	106.1	3.0	0.9
Middle quartiles	6.6	24.9	32.5	42.7	158.7	3.5	0.5
Upper quartile	9.2	22.0	29.9	40.9	198.8	3.4	0.4
Overall	2.9	22.2	31.0	41.2	198.8	3.4	0.6

10.2.2 SP71115 (TIMSS)

10.2.2.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	56.71%	299	TIMSS	MC	0.26	0.25	0.57	−0.47	0.017

10.2.2.2 Gender DIF:

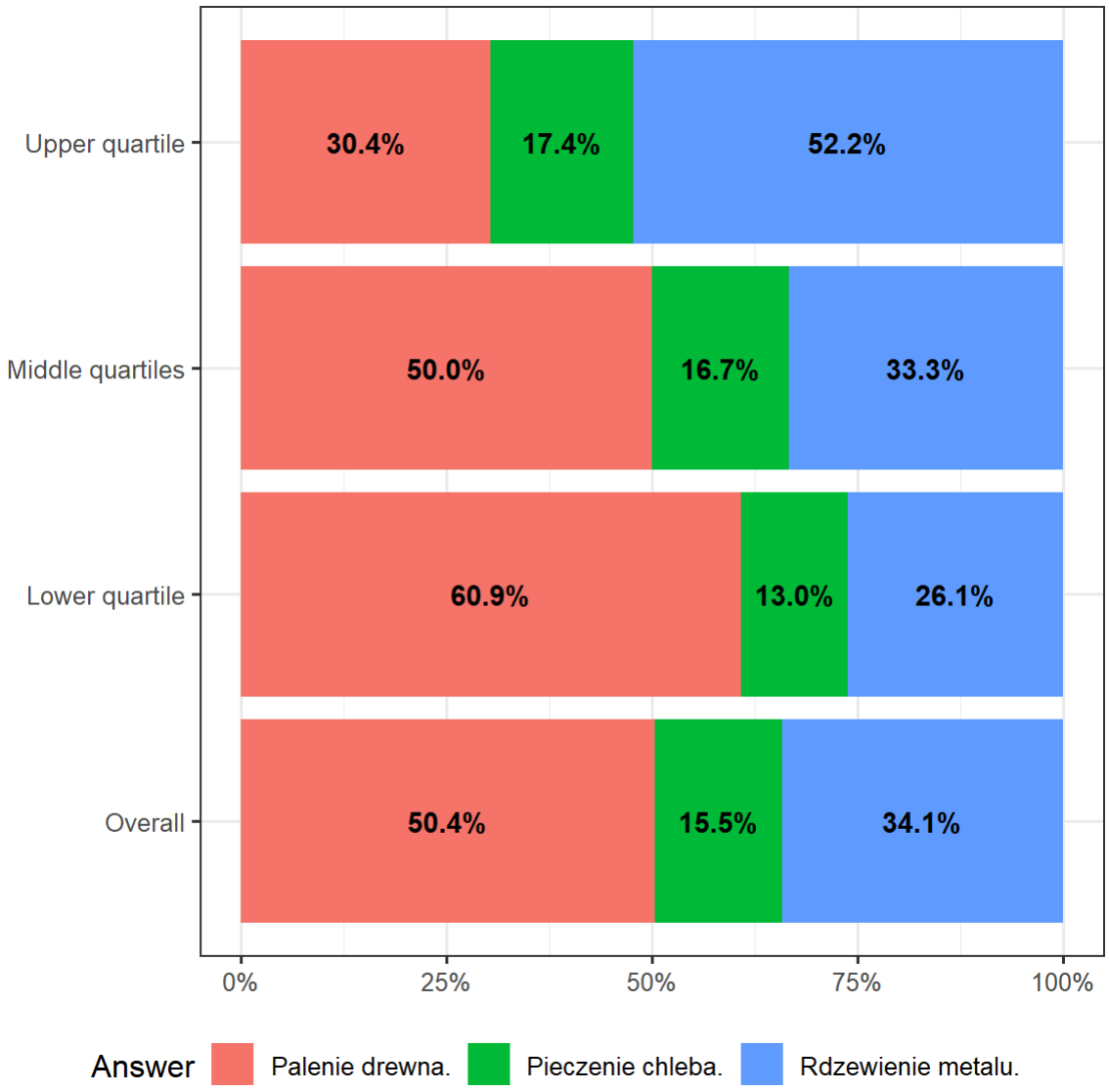
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.85	0.397	1.000	0.005	N/A

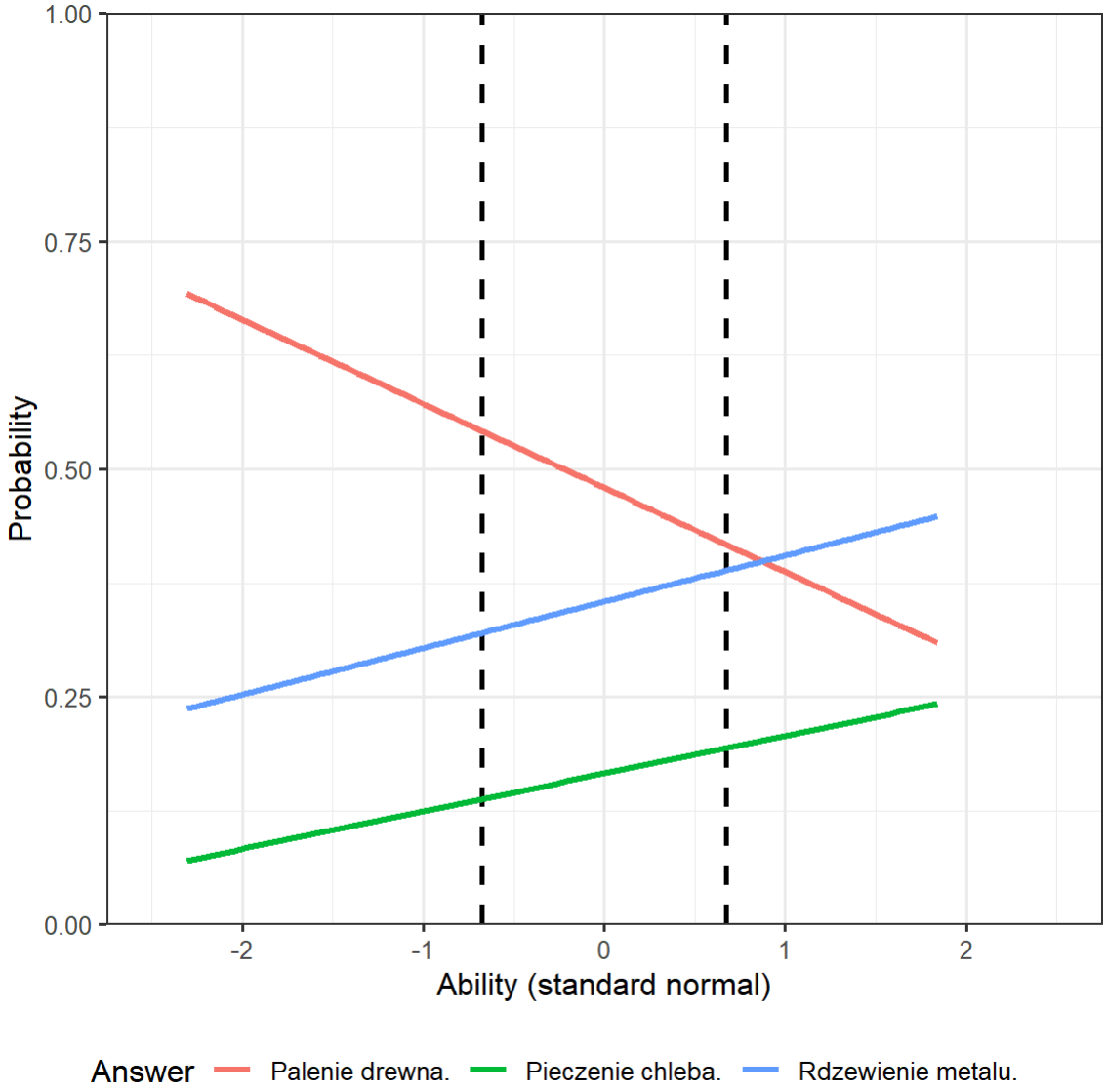
10.2.2.3 Frequency of answers:

Answer	Frequency	Percent
Palenie drewna.	65	21.8%
Pieczenie chleba.	20	6.7%
Rdzewienie metalu.	44	14.8%
Rozciąganie gumy. (correct)	169	56.7%
Overall	298	100.0%

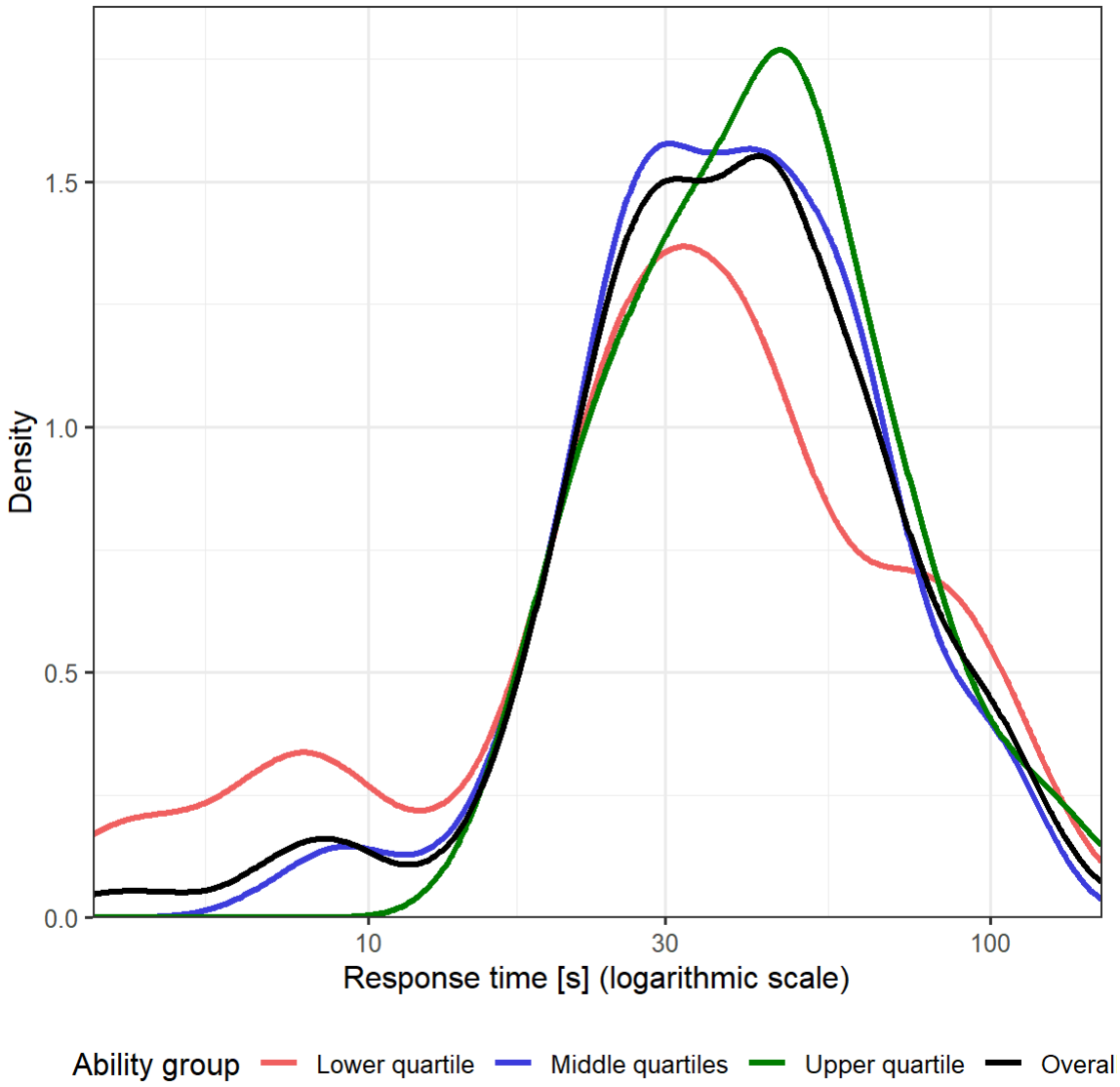
10.2.2.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Palenie drewna.	28	30	7	65	60.9%	50.0%	30.4%	50.4%
Pieczenie chleba.	6	10	4	20	13.0%	16.7%	17.4%	15.5%
Rdzewienie metalu.	12	20	12	44	26.1%	33.3%	52.2%	34.1%
Overall	46	60	23	129	100.0%	100.0%	100.0%	100.0%


Upper quartile
Middle quartiles
Lower quartile
Overall
30.4%
17.4%
52.2%
50.0%
16.7%
33.3%
60.9%
13.0%
26.1%
50.4%
15.5%
34.1%
0%
25%
50%
75%
100%
Answer
Palenie drewna.
Pieczenie chleba.
Rdzewienie metalu.



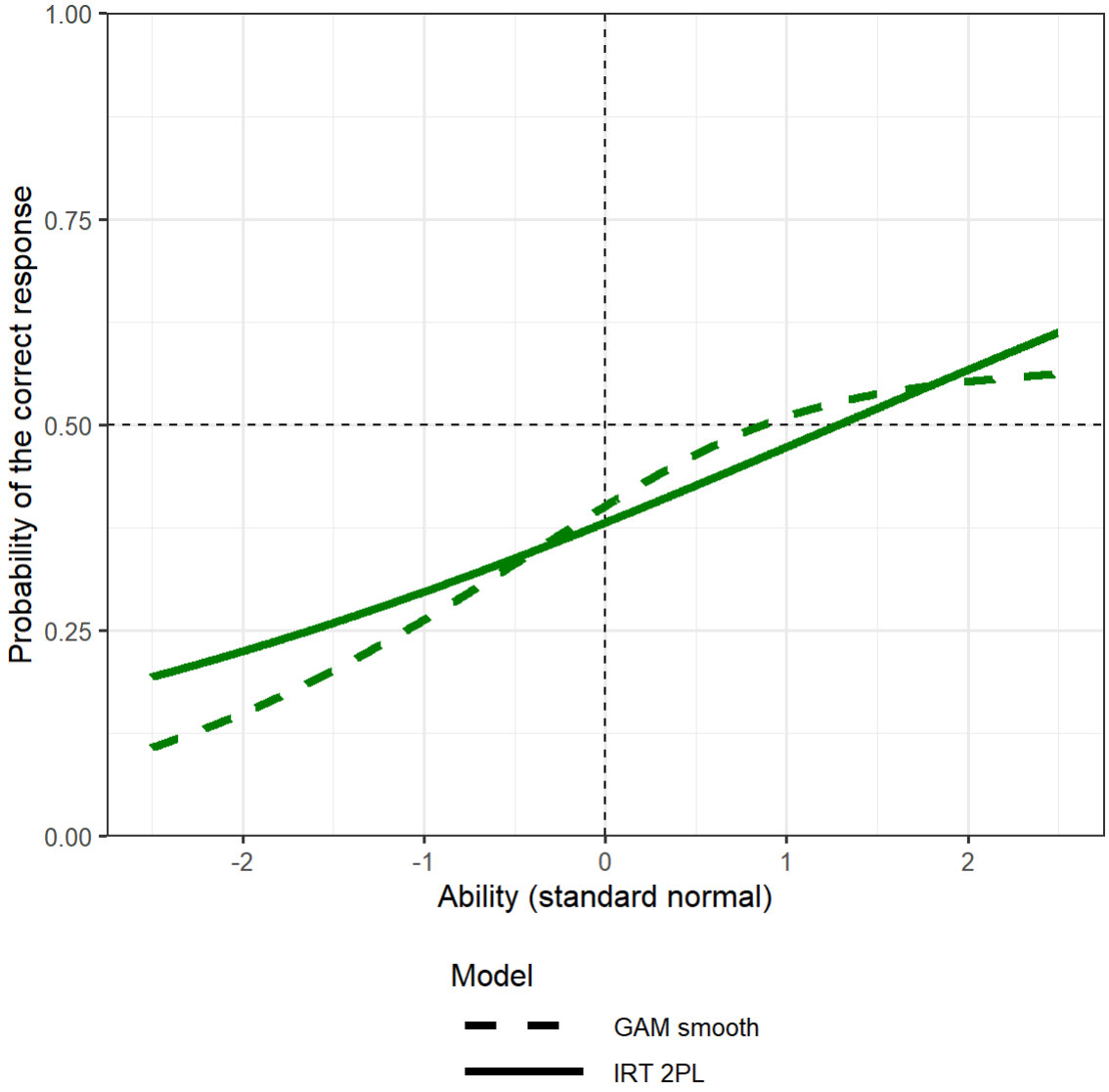
10.2.2.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.6	21.7	32.5	47.7	133.6	3.4	0.8
Middle quartiles	7.1	26.4	38.5	55.8	114.4	3.6	0.5
Upper quartile	15.6	29.6	44.4	57.1	150.6	3.7	0.5
Overall	3.6	26.1	38.2	55.8	150.6	3.6	0.6

10.2.3 S061098 (TIMSS)

10.2.3.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	38.03%	306	TIMSS	MC	0.23	0.12	0.38	1.29	0.008

10.2.3.2 Gender DIF:

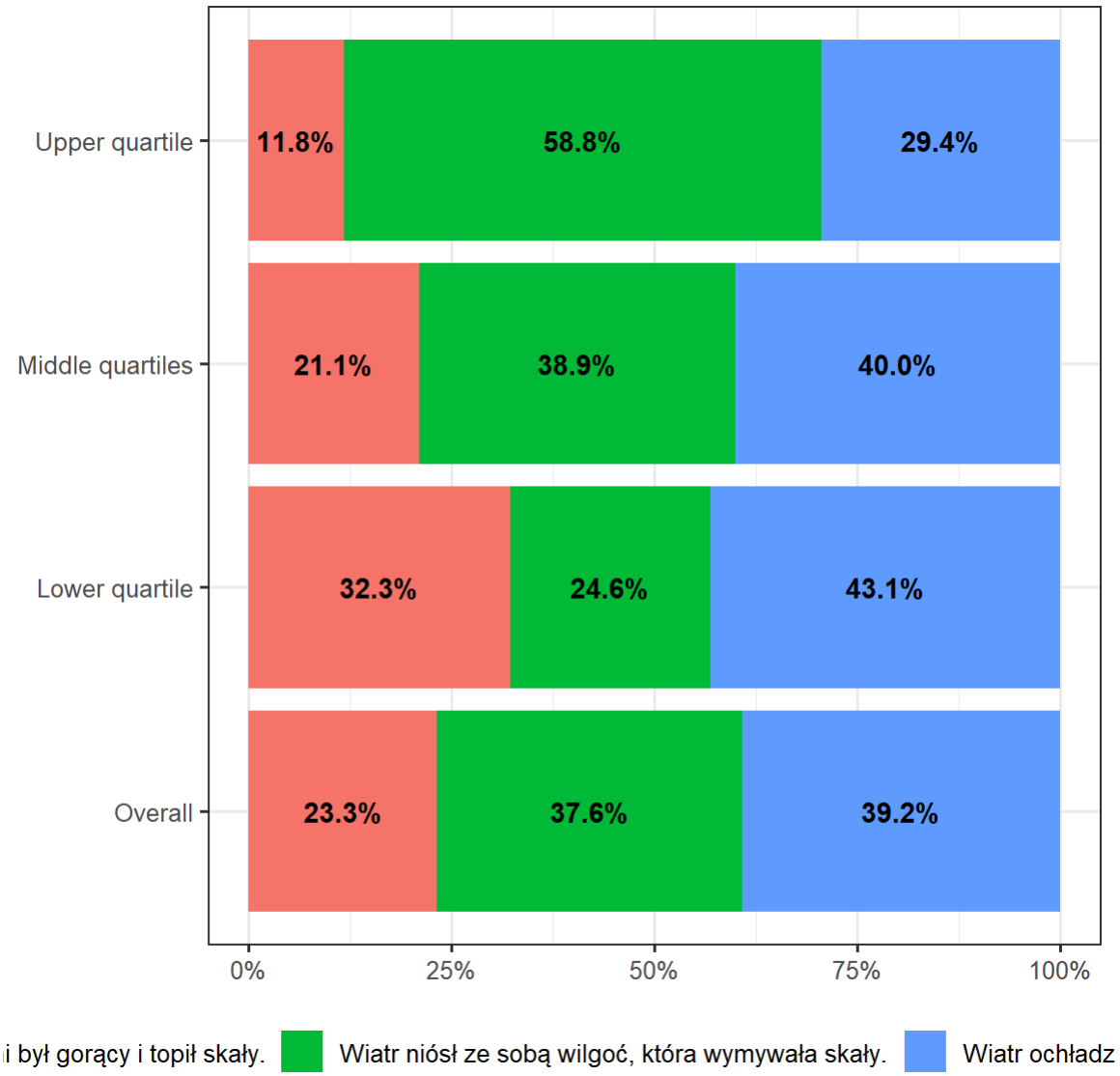
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.17	0.557	1.000	0.003	N/A

10.2.3.3 Frequency of answers:

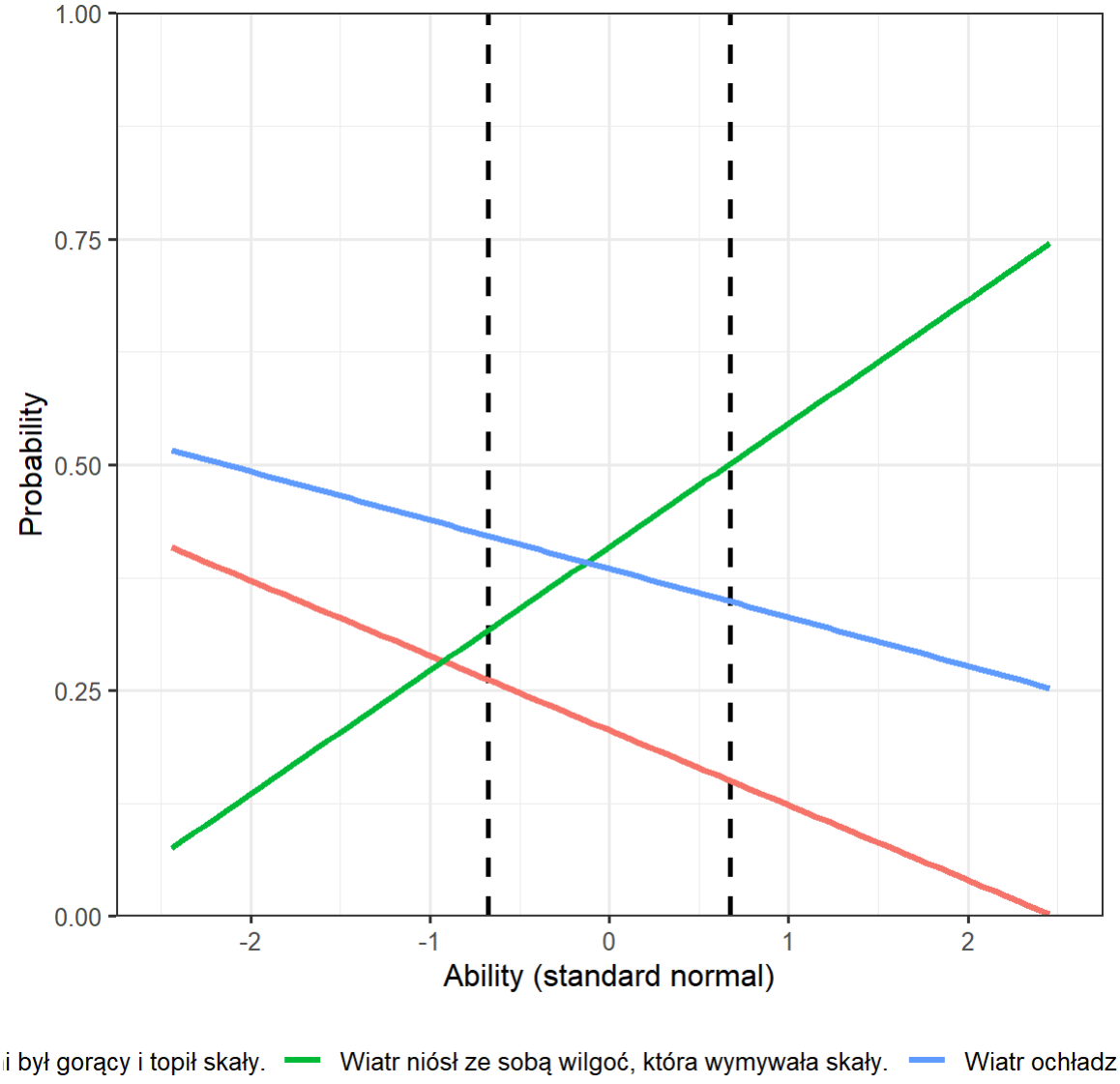
Answer	Frequency	Percent
Wiatr na pustyni był gorący i topił skały.	44	14.4%
Wiatr niósł ze sobą wilgoć, która wymywała skały.	71	23.3%
Wiatr niósł ze sobą ziarenka piasku, które tarły o skałę. (correct)	116	38.0%
Wiatr ochładzał skały, które następnie pękały.	74	24.3%
Overall	305	100.0%

10.2.3.4 Distractors functioning:

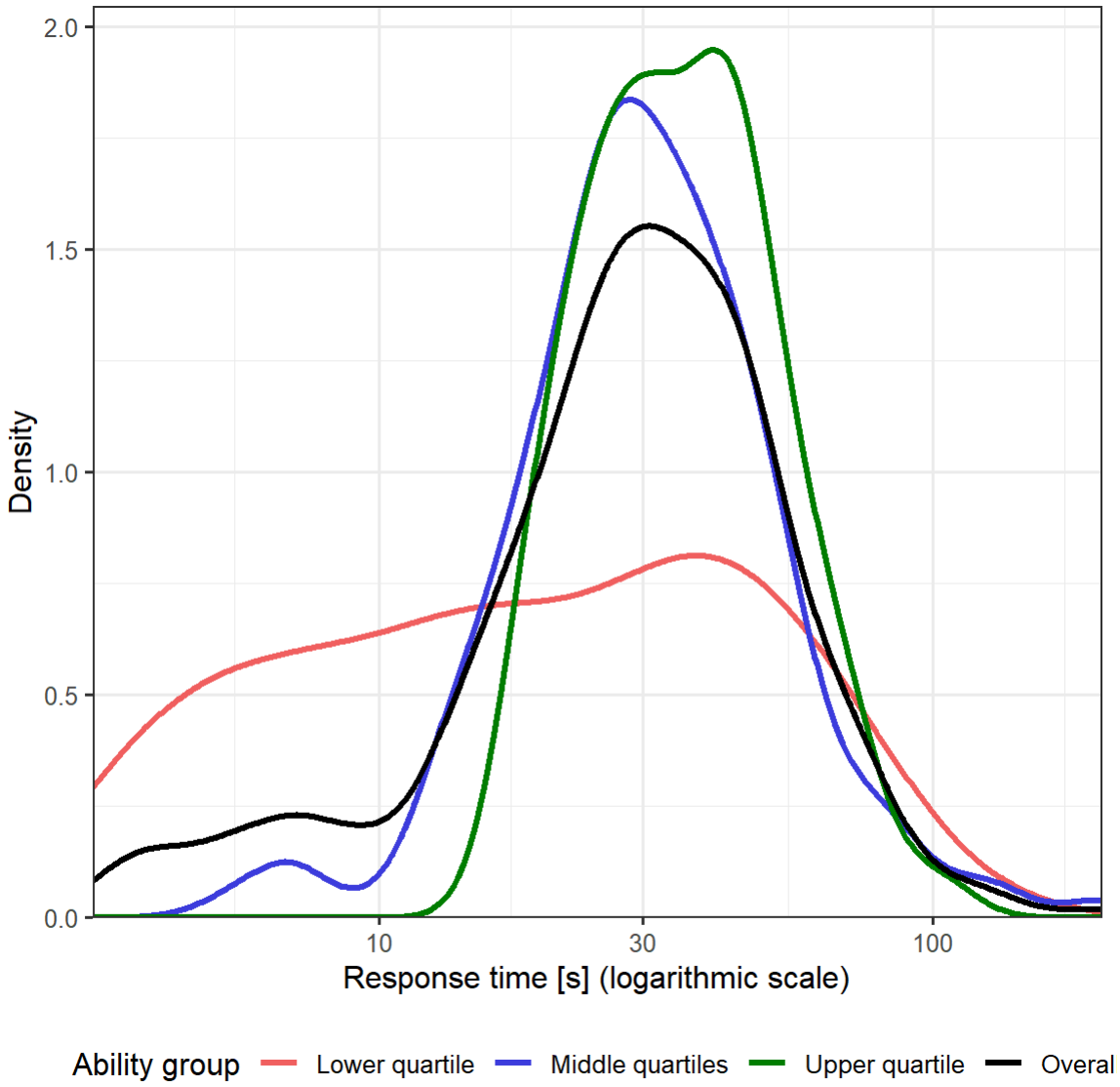
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Wiatr na pustyni był gorący i topił skały.	21	19	4	44	32.3%	21.1%	11.8%	23.3%
Wiatr niósł ze sobą wilgoć, która wymywała skały.	16	35	20	71	24.6%	38.9%	58.8%	37.6%
Wiatr ochładzał skały, które następnie pękały.	28	36	10	74	43.1%	40.0%	29.4%	39.2%
Overall	65	90	34	189	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 9 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



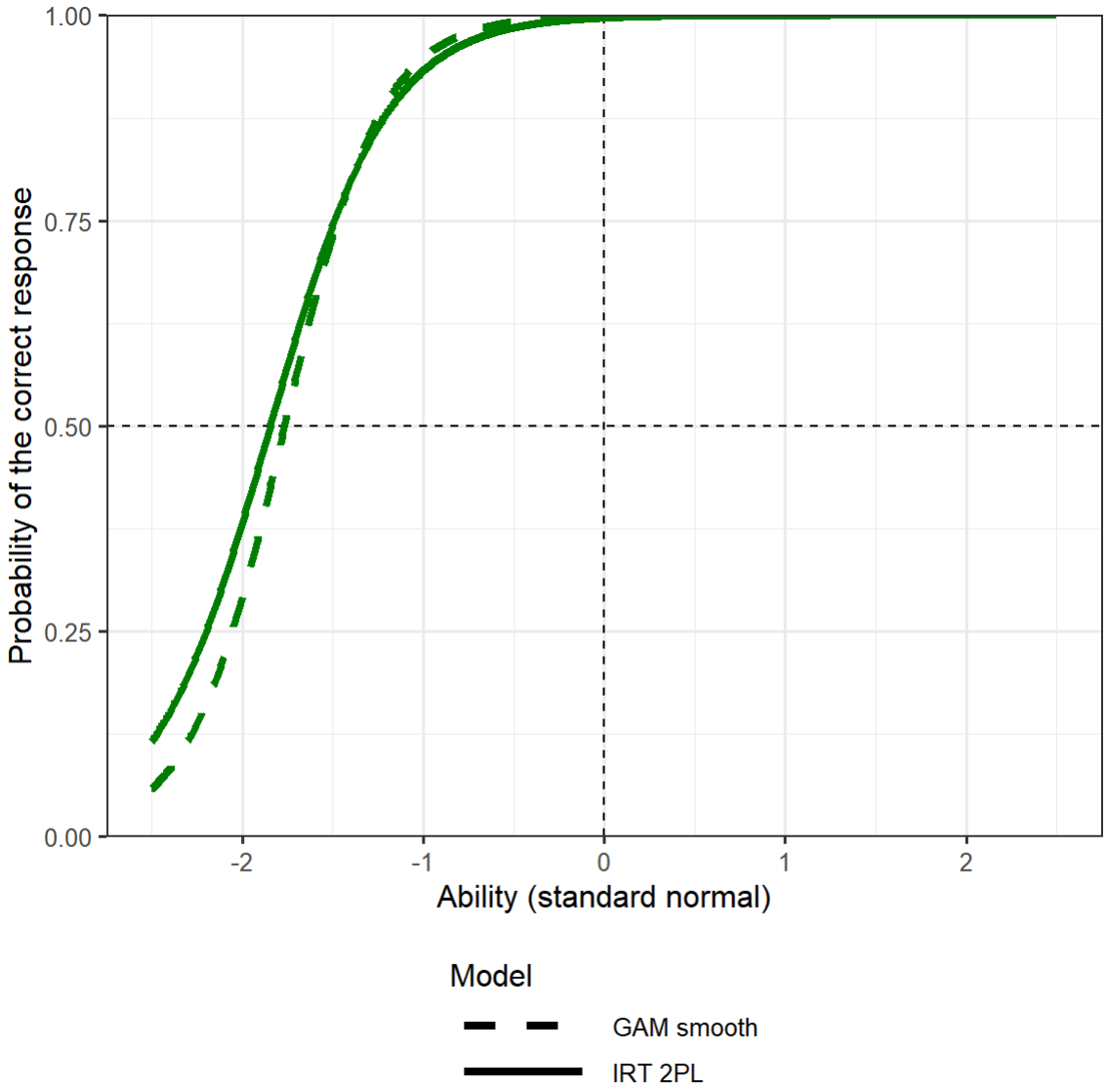
10.2.3.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	8.1	18.0	39.2	117.0	2.9	1.0
Middle quartiles	5.4	22.3	29.7	42.6	201.7	3.4	0.5
Upper quartile	17.9	25.8	34.1	44.4	102.0	3.6	0.4
Overall	3.0	19.5	29.7	43.0	201.7	3.3	0.7

10.2.4 S090052 (LLM)

10.2.4.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	94.63%	299	LLM teachers' selection	MC	0.45	0.28	3.10	−1.84	0.0(

10.2.4.2 Gender DIF:

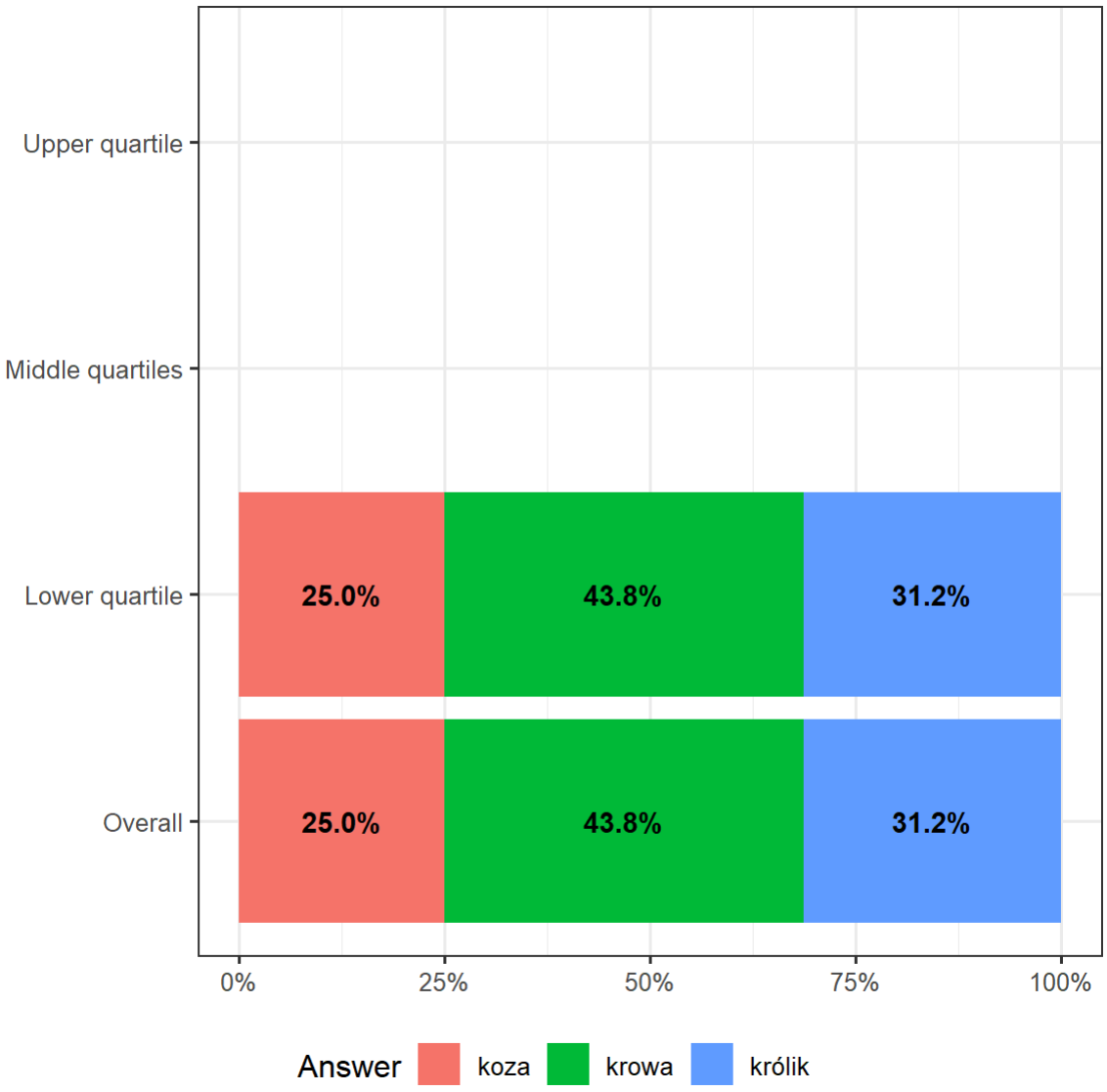
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.76	0.153	1.000	0.031	N/A

10.2.4.3 Frequency of answers:

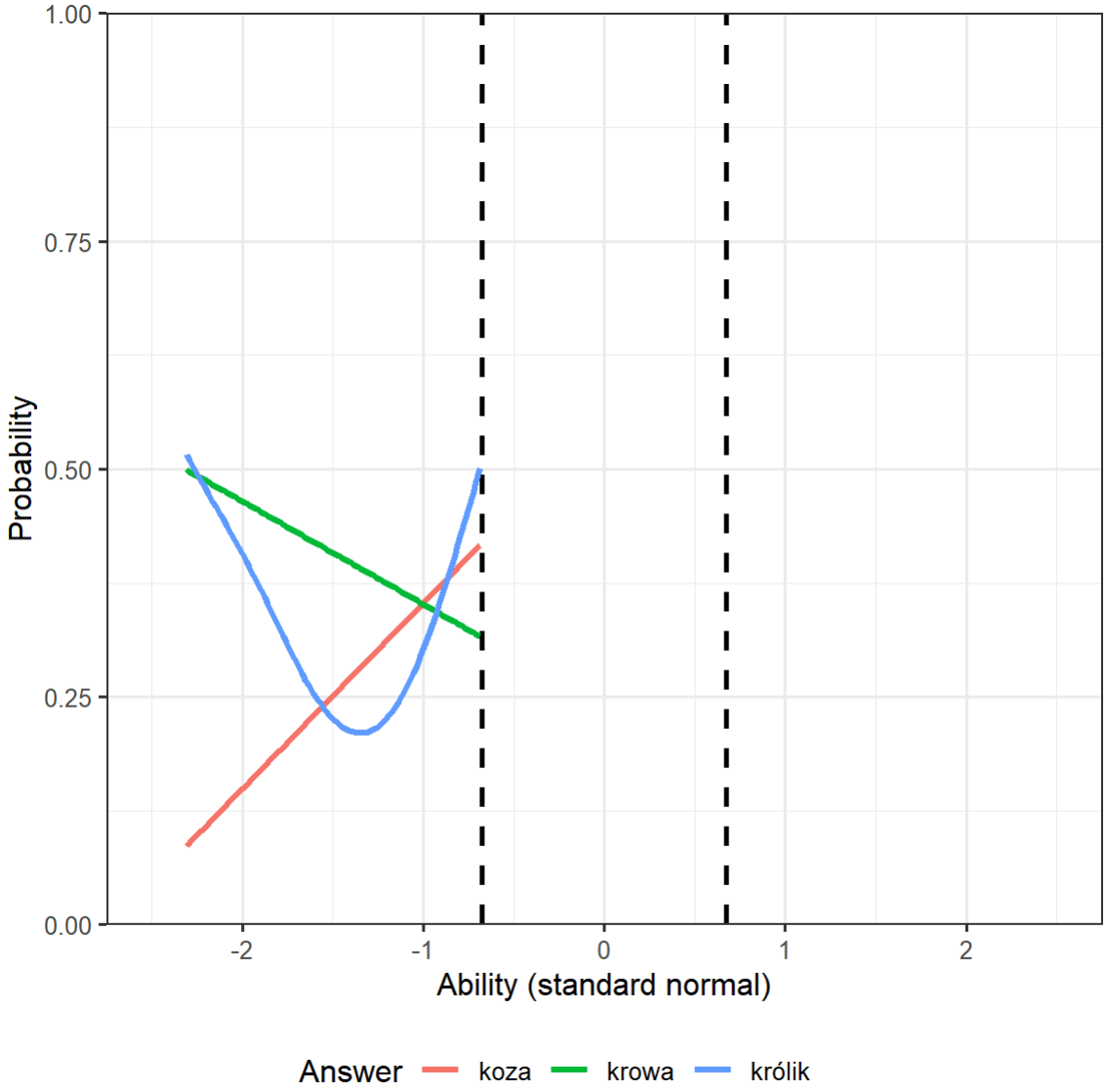
Answer	Frequency	Percent
koza	4	1.3%
krowa	7	2.3%
królik	5	1.7%
lew (correct)	282	94.6%
Overall	298	100.0%

10.2.4.4 Distractors functioning:

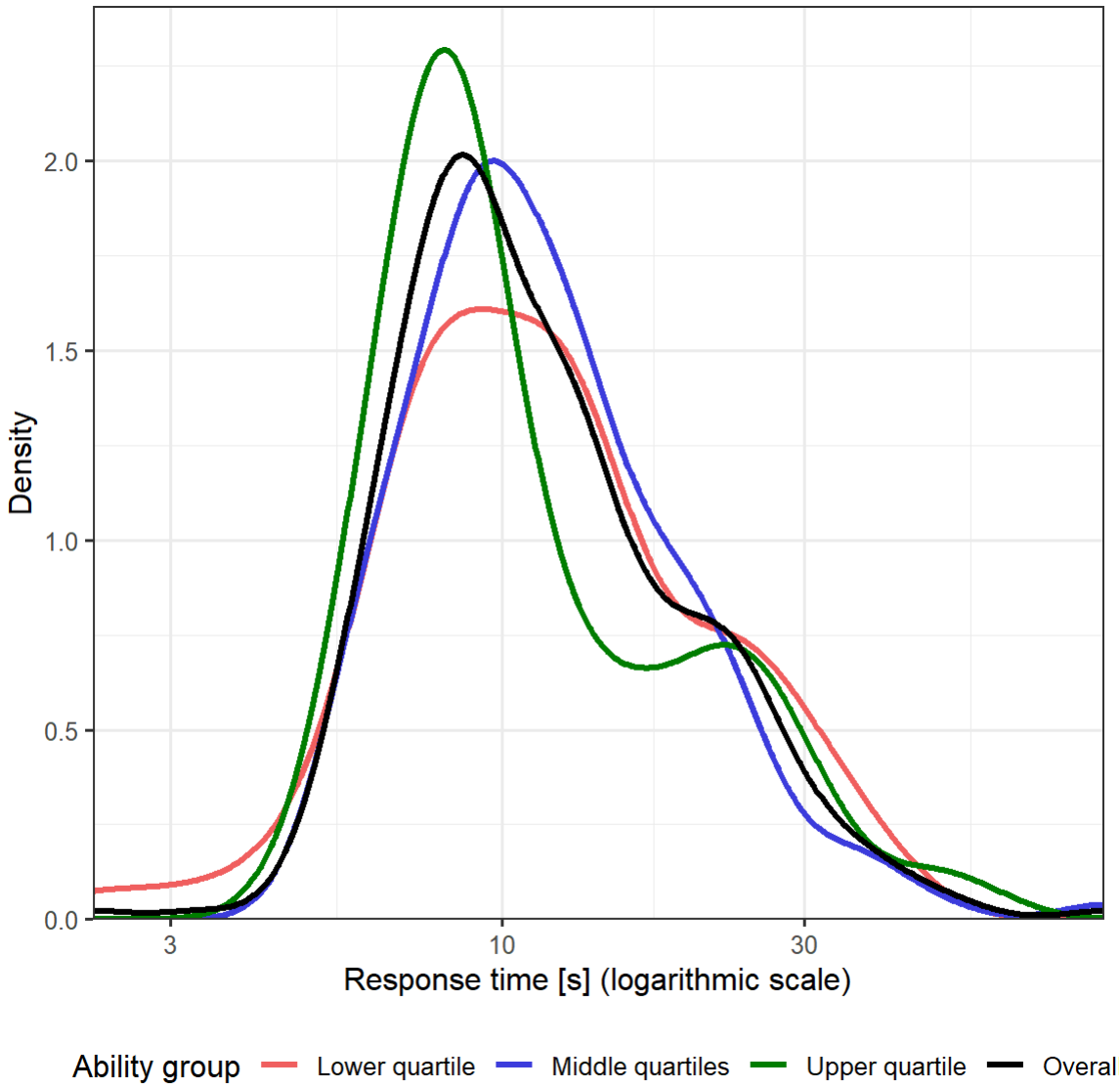
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
koza	4	0	0	4	25.0%	0.0%	0.0%	25.0%
krowa	7	0	0	7	43.8%	0.0%	0.0%	43.8%
królik	5	0	0	5	31.2%	0.0%	0.0%	31.2%
Overall	16	0	0	16	100.0%	0.0%	0.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



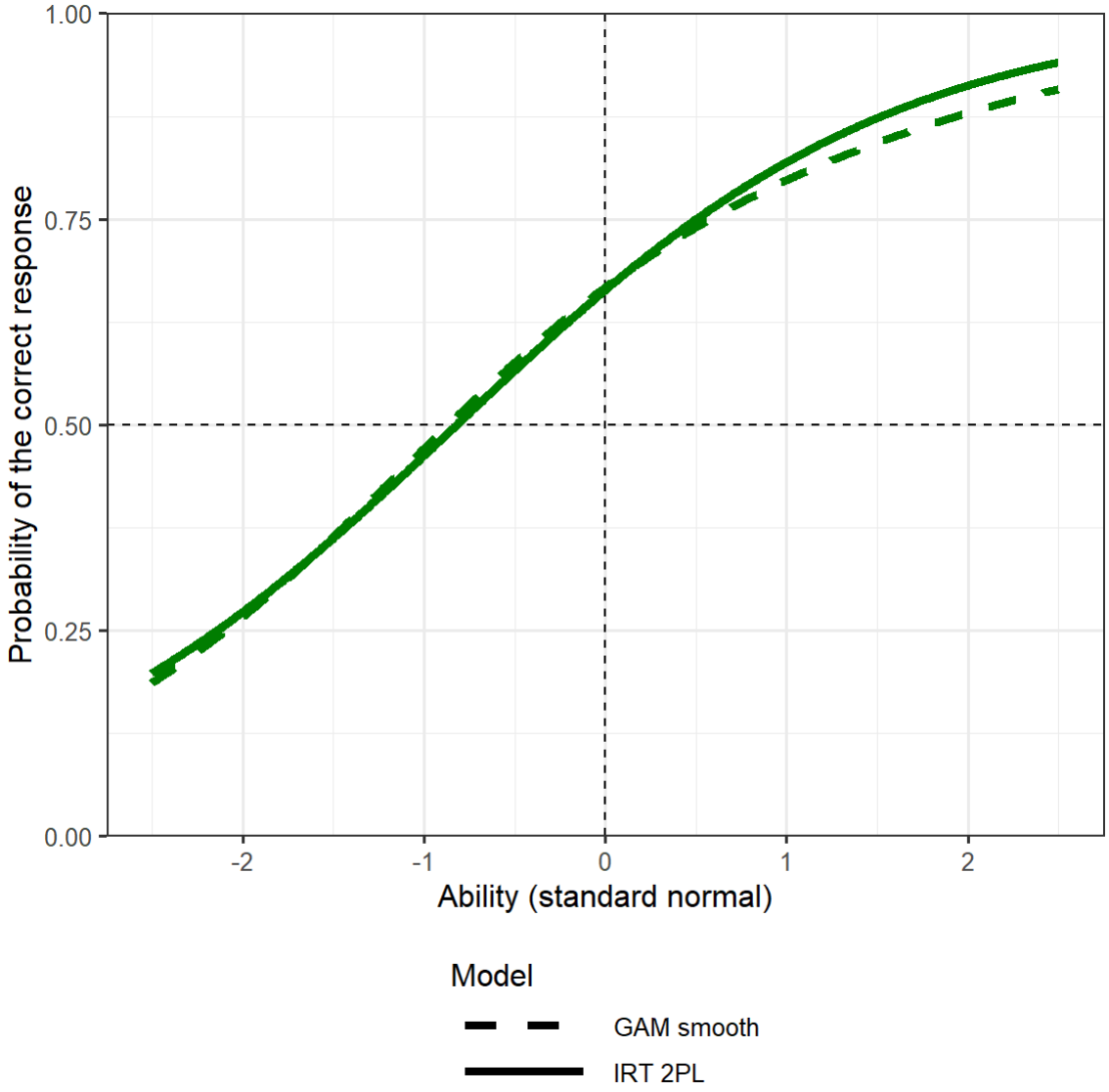
10.2.4.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.3	7.8	11.5	16.1	38.9	2.4	0.6
Middle quartiles	5.0	8.4	10.8	15.7	88.8	2.5	0.5
Upper quartile	5.4	7.6	8.8	15.3	53.5	2.4	0.5
Overall	2.3	8.0	10.3	15.7	88.8	2.4	0.5

10.2.5 S061065 (LLM)

10.2.5.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	63.82%	305	TIMSS	MC	0.31	0.35	0.83	−0.82	0.015

10.2.5.2 Gender DIF:

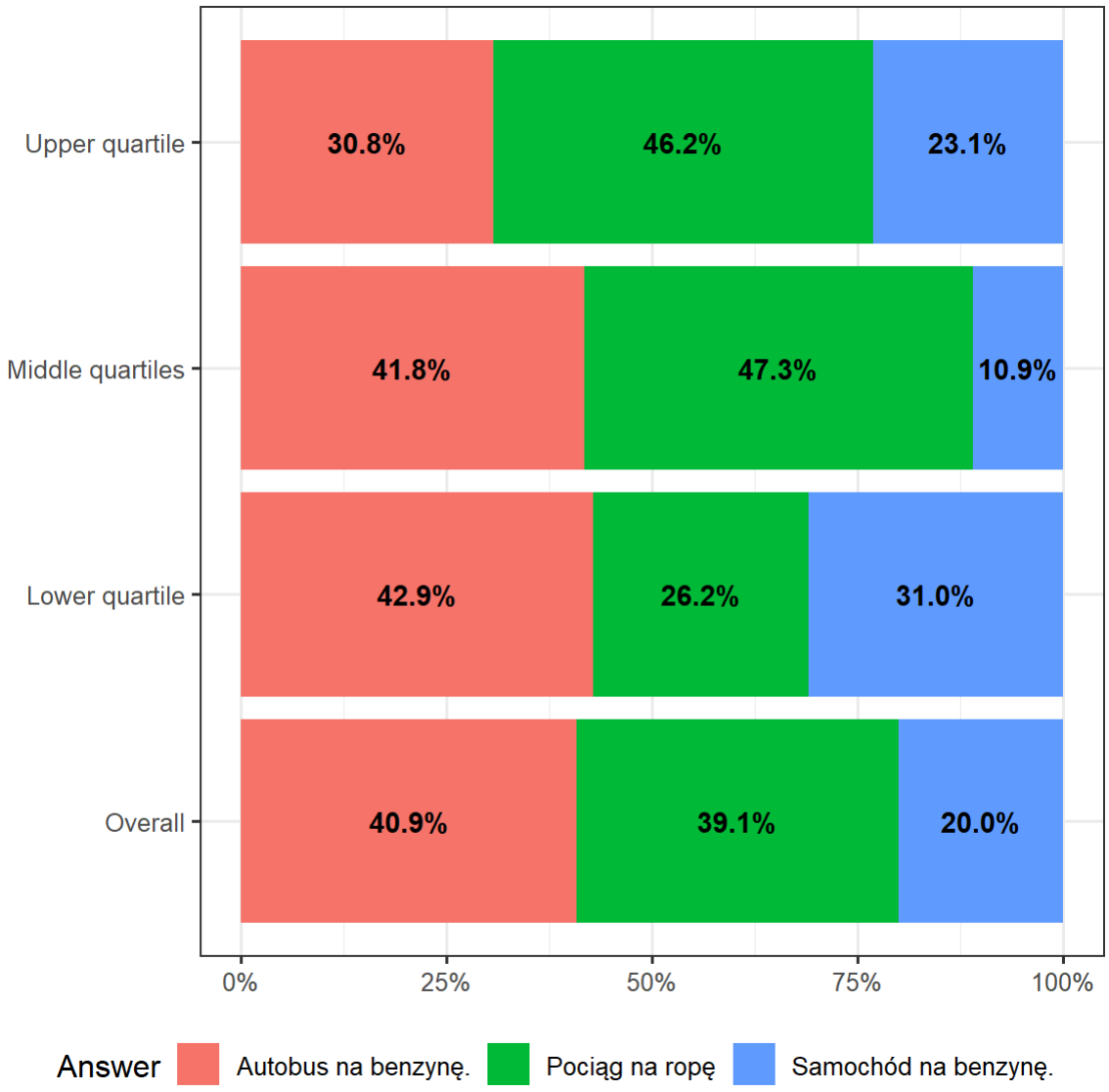
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
4.80	0.091	1.000	0.014	N/A

10.2.5.3 Frequency of answers:

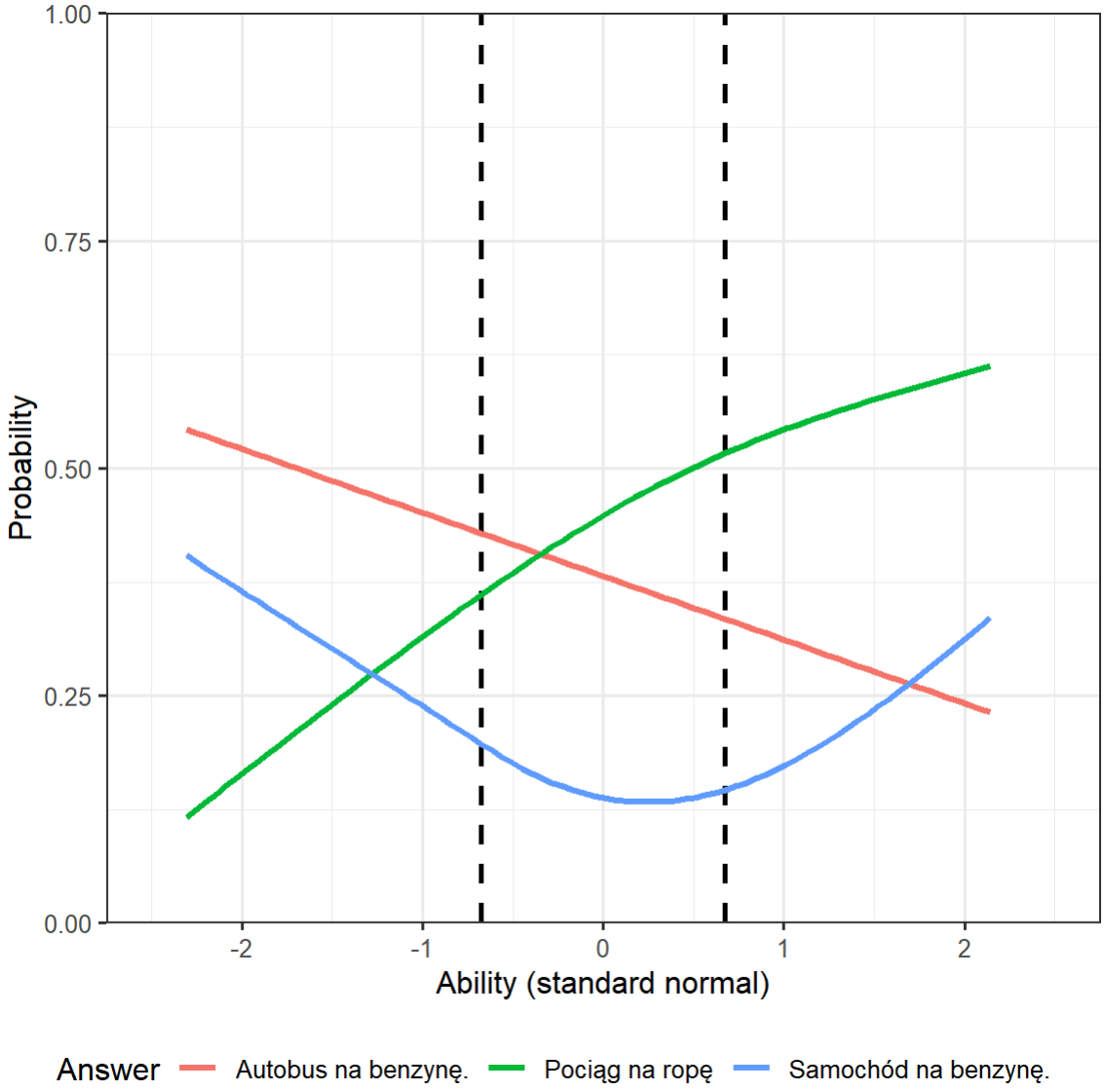
Answer	Frequency	Percent
Autobus na benzynę.	45	14.8%
Pociąg na ropę	43	14.1%
Samochód elektryczny. (correct)	194	63.8%
Samochód na benzynę.	22	7.2%
Overall	304	100.0%

10.2.5.4 Distractors functioning:

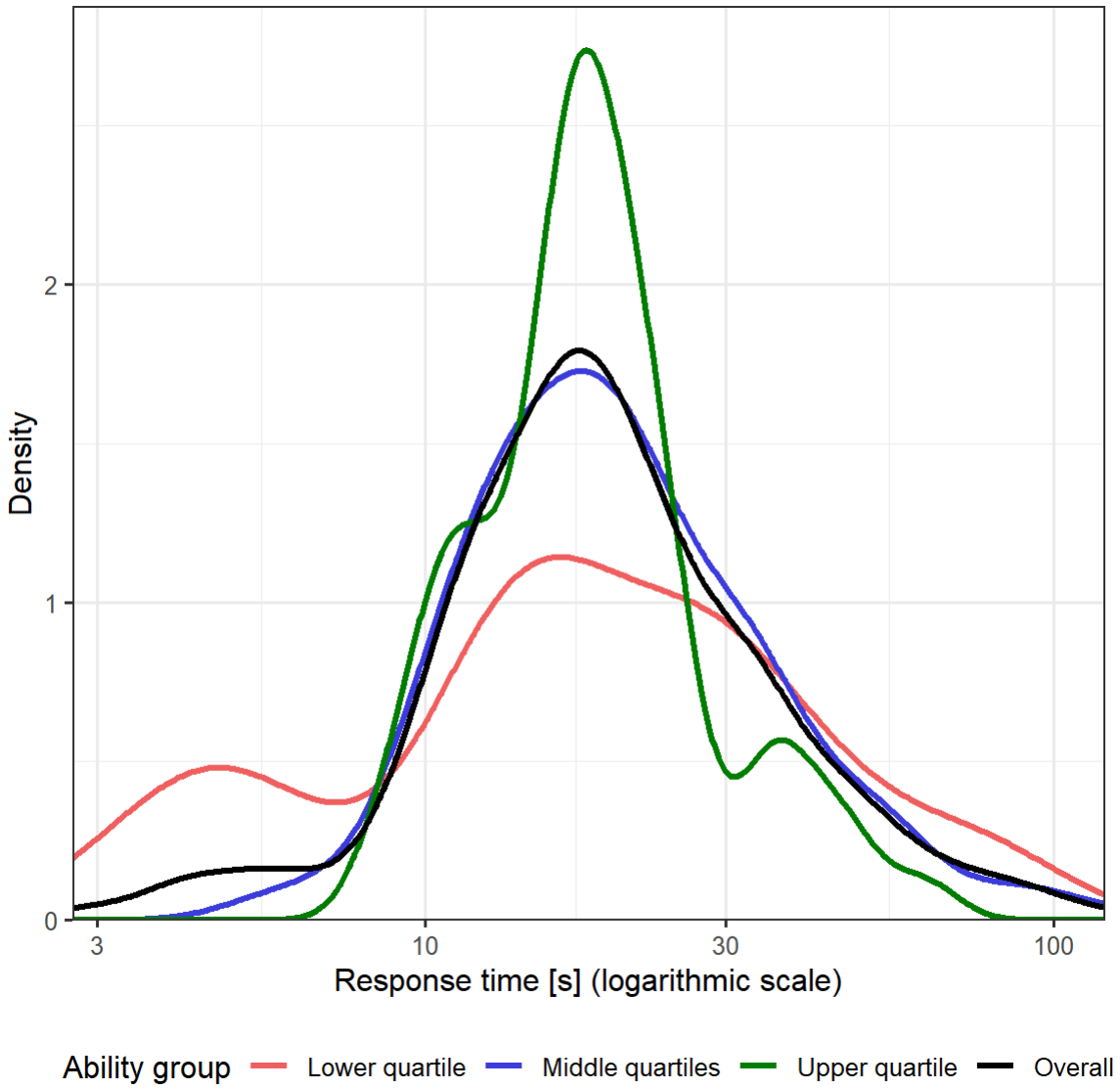
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Autobus na benzynę.	18	23	4	45	42.9%	41.8%	30.8%	40.9%
Pociąg na ropę	11	26	6	43	26.2%	47.3%	46.2%	39.1%
Samochód na benzynę.	13	6	3	22	31.0%	10.9%	23.1%	20.0%
Overall	42	55	13	110	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



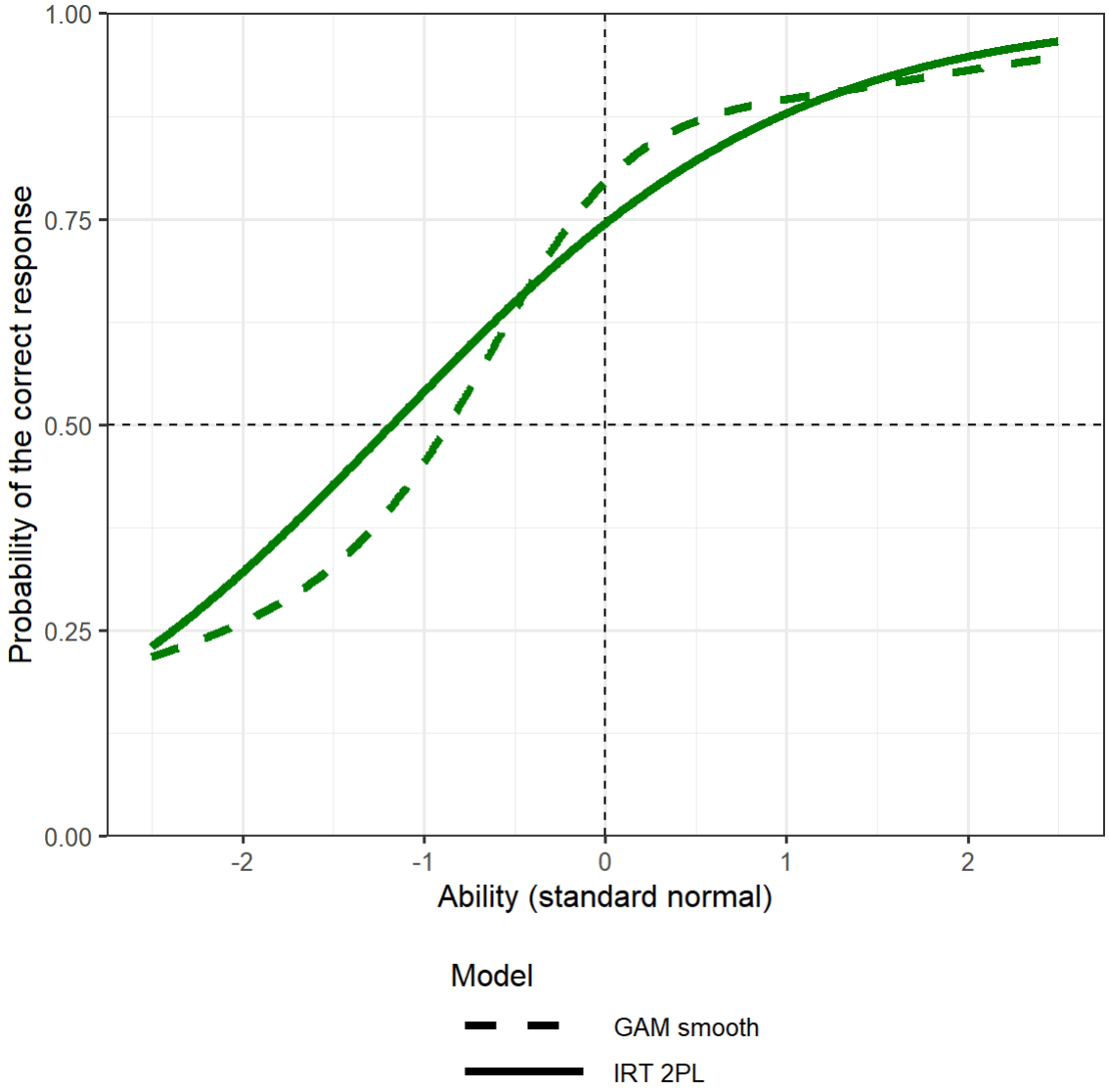
10.2.5.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	11.8	17.3	30.6	100.7	2.8	0.9
Middle quartiles	5.3	13.6	19.1	28.0	120.3	3.0	0.6
Upper quartile	8.7	14.6	17.8	22.2	62.2	2.9	0.4
Overall	2.7	13.3	18.4	27.1	120.3	3.0	0.6

10.2.6 S100002 (TIMSS)

10.2.6.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	71.81%	299	LLM simple setup	MC	0.42	0.31	0.91	−1.18	0.022

10.2.6.2 Gender DIF:

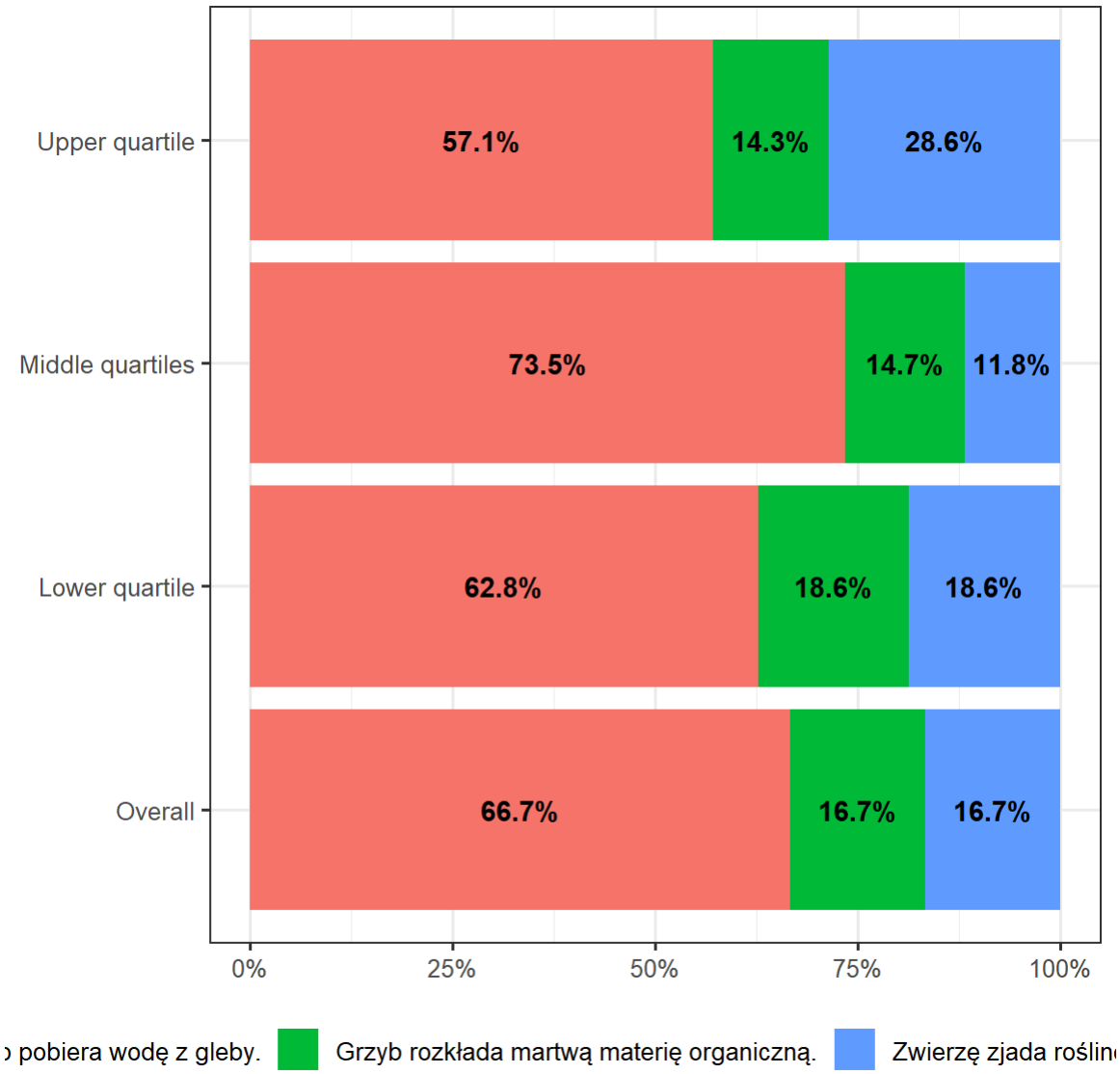
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.53	0.766	1.000	0.002	N/A

10.2.6.3 Frequency of answers:

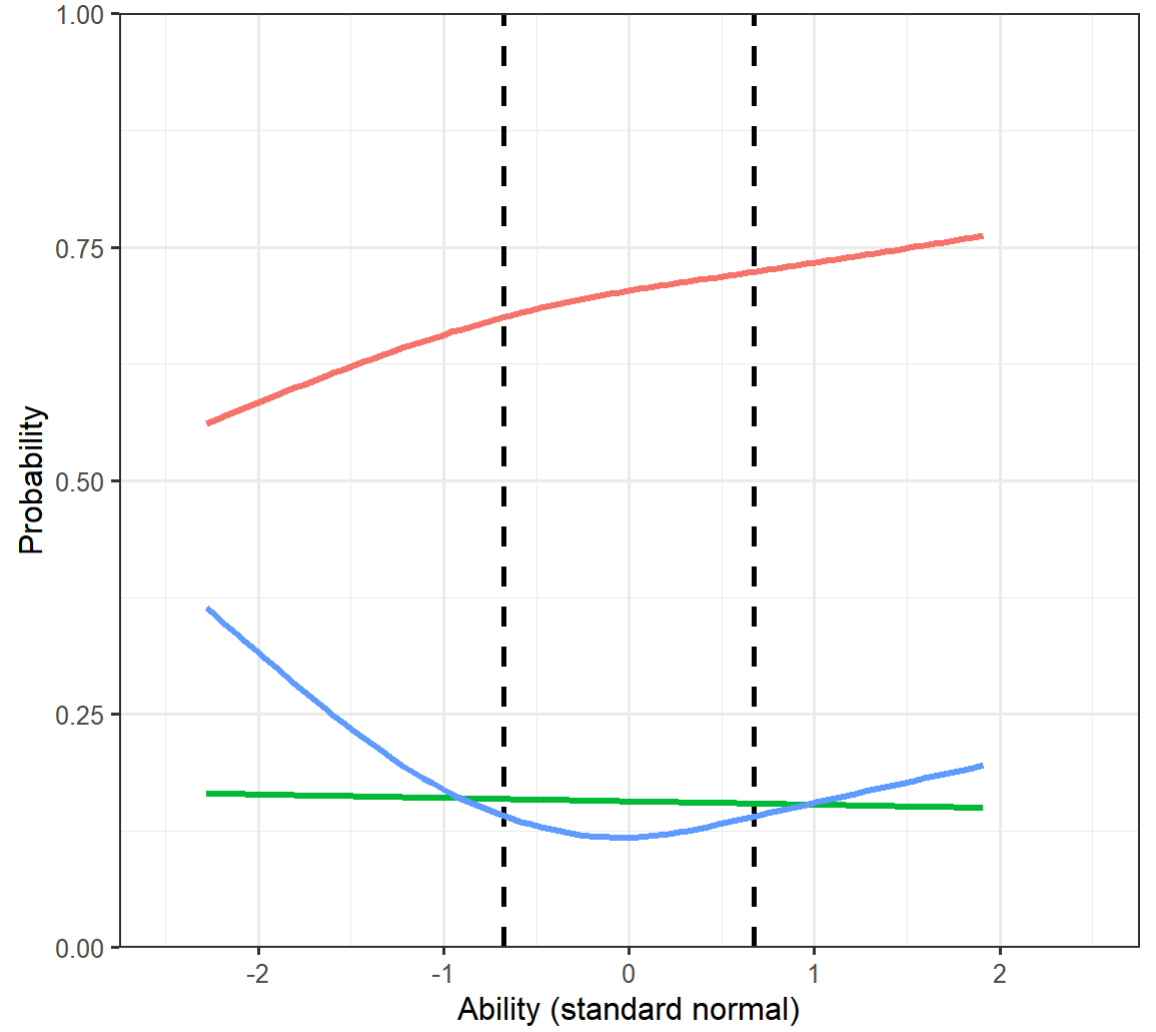
Answer	Frequency	Percent
Drzewo pobiera wodę z gleby.	56	18.8%
Grzyb rozkłada martwą materię organiczną.	14	4.7%
Roślina przekształca dwutlenek węgla i wodę w glukozę i tlen przy użyciu światła słonecznego. (correct)	214	71.8%
Zwierzę zjada roślinę, aby uzyskać energię.	14	4.7%
Overall	298	100.0%

10.2.6.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Drzewo pobiera wodę z gleby.	27	25	4	56	62.8%	73.5%	57.1%	66.7%
Grzyb rozkłada martwą materię organiczną.	8	5	1	14	18.6%	14.7%	14.3%	16.7%
Zwierzę zjada roślinę, aby uzyskać energię.	8	4	2	14	18.6%	11.8%	28.6%	16.7%
Overall	43	34	7	84	100.0%	100.0%	100.0%	100.0%

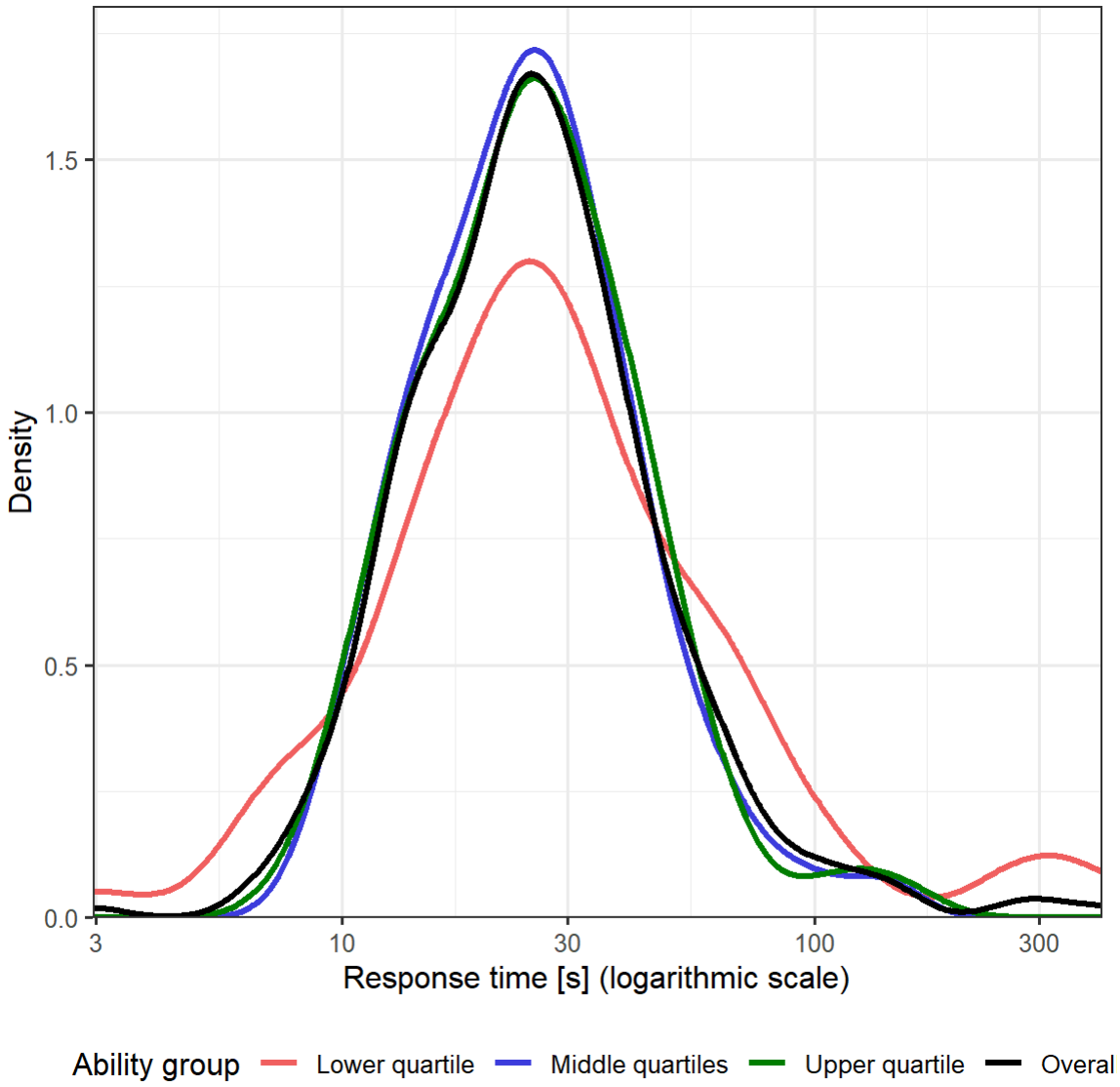



```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



› pobiera wodę z gleby. Grzyb rozkłada martwą materię organiczną. Zwierzę zjada roślin

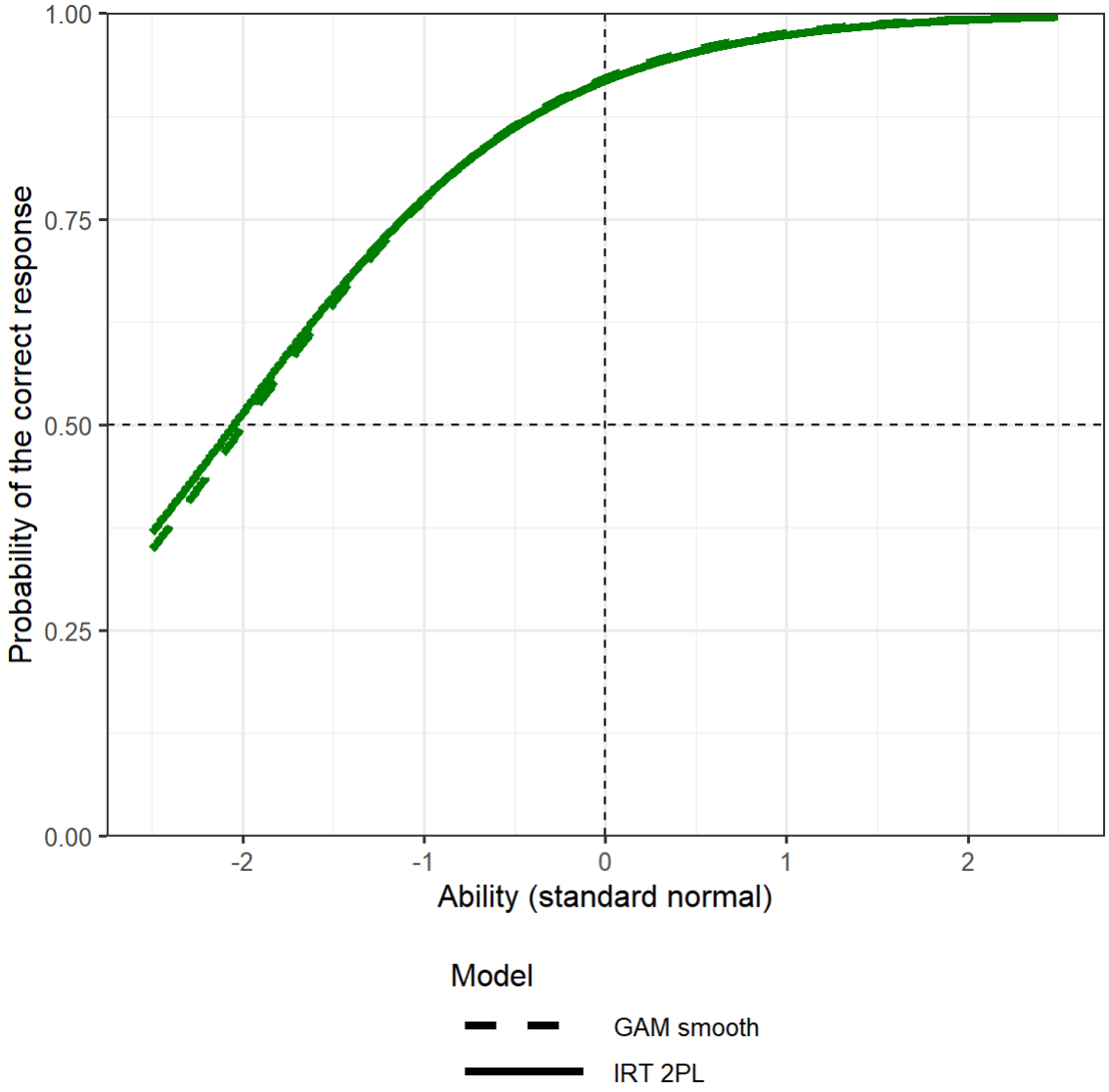
10.2.6.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	17.5	26.1	43.4	405.7	3.3	0.9
Middle quartiles	8.7	17.0	24.6	34.2	143.7	3.2	0.5
Upper quartile	8.3	17.1	24.5	35.9	147.6	3.2	0.5
Overall	3.0	17.0	24.7	36.1	405.7	3.2	0.6

10.2.7 S090040 (TIMSS)

10.2.7.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	87.92%	299	LLM teachers' selection	MC	0.34	0.17	1.18	−2.05	0.0(

10.2.7.2 Gender DIF:

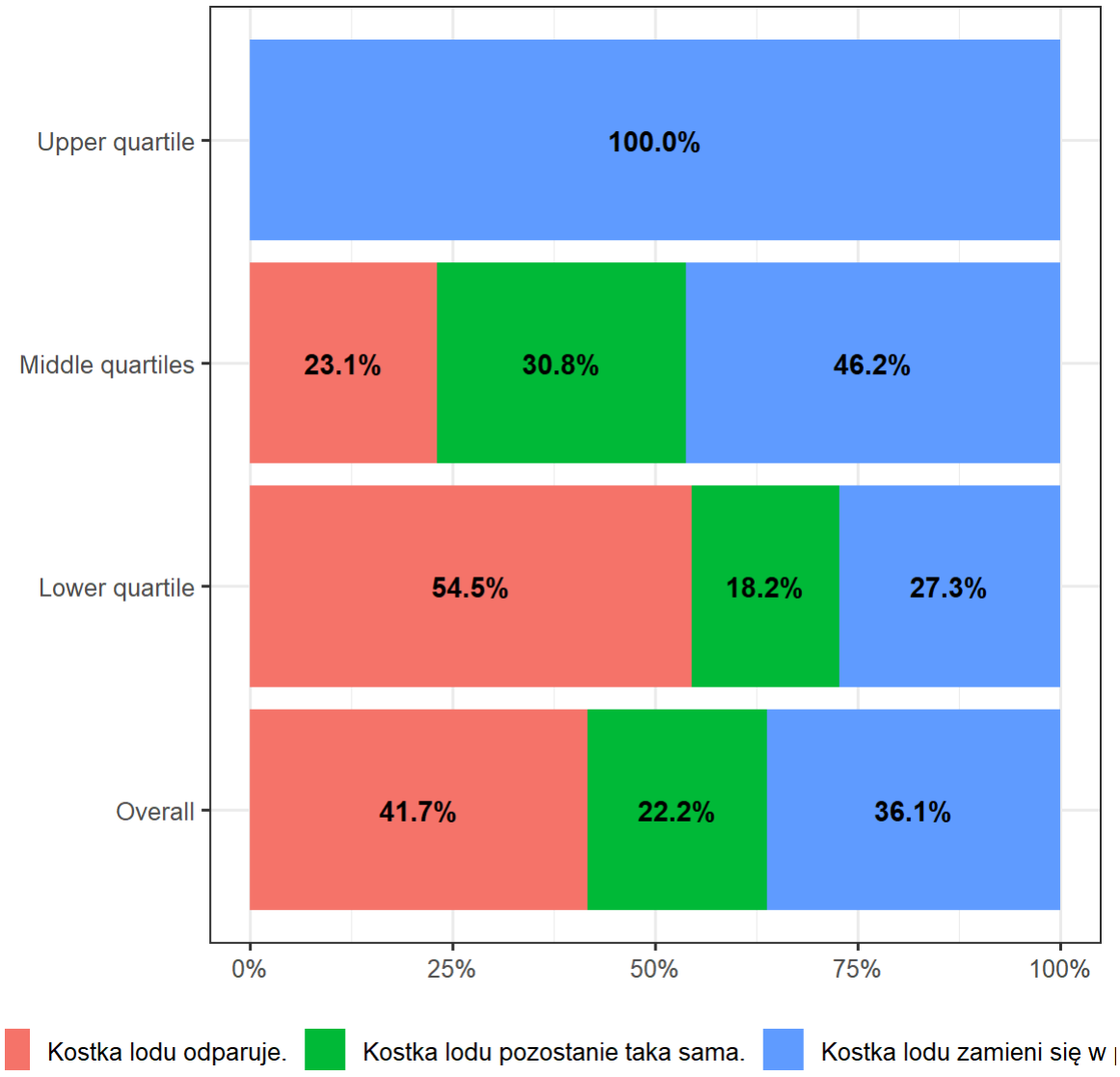
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.04	0.596	1.000	0.005	N/A

10.2.7.3 Frequency of answers:

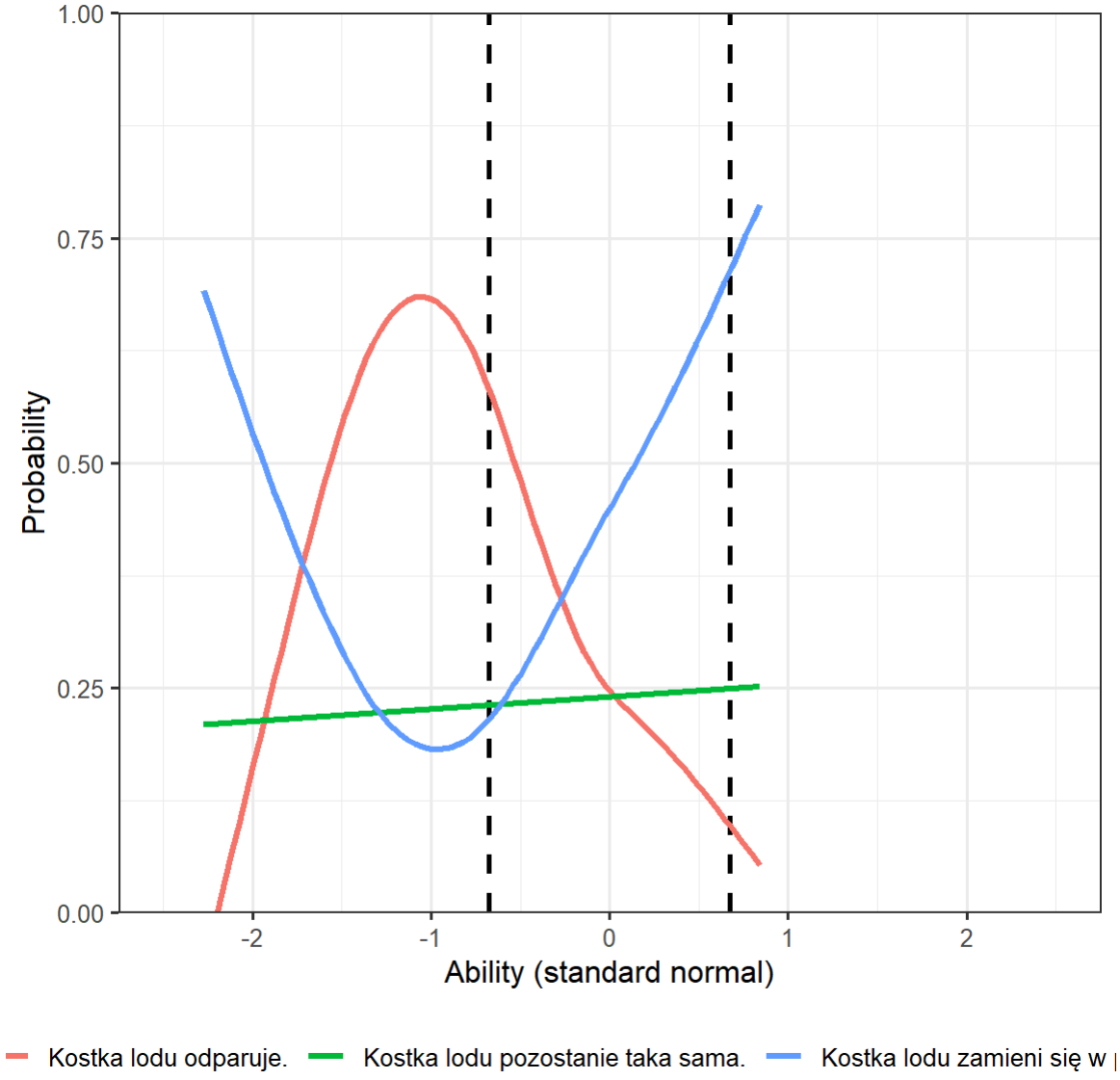
Answer	Frequency	Percent
Kostka lodu odparuje.	15	5.0%
Kostka lodu pozostanie taka sama.	8	2.7%
Kostka lodu zamieni się w parę wodną.	13	4.4%
Kostka lodu zmieni się w wodę. (correct)	262	87.9%
Overall	298	100.0%

10.2.7.4 Distractors functioning:

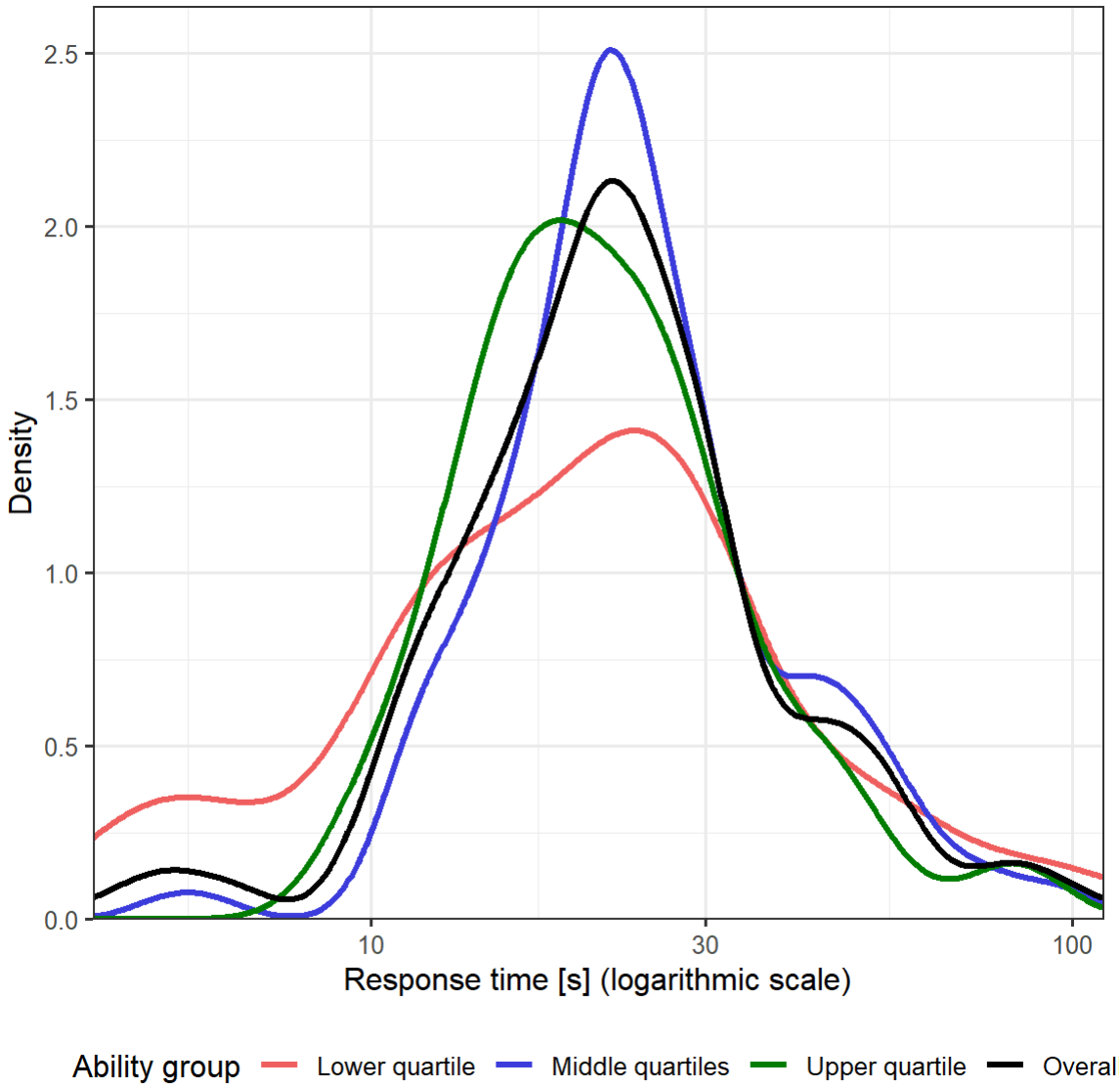
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Kostka lodu odparuje.	12	3	0	15	54.5%	23.1%	0.0%	41.7%
Kostka lodu pozostanie taka sama.	4	4	0	8	18.2%	30.8%	0.0%	22.2%
Kostka lodu zamieni się w parę wodną.	6	6	1	13	27.3%	46.2%	100.0%	36.1%
Overall	22	13	1	36	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



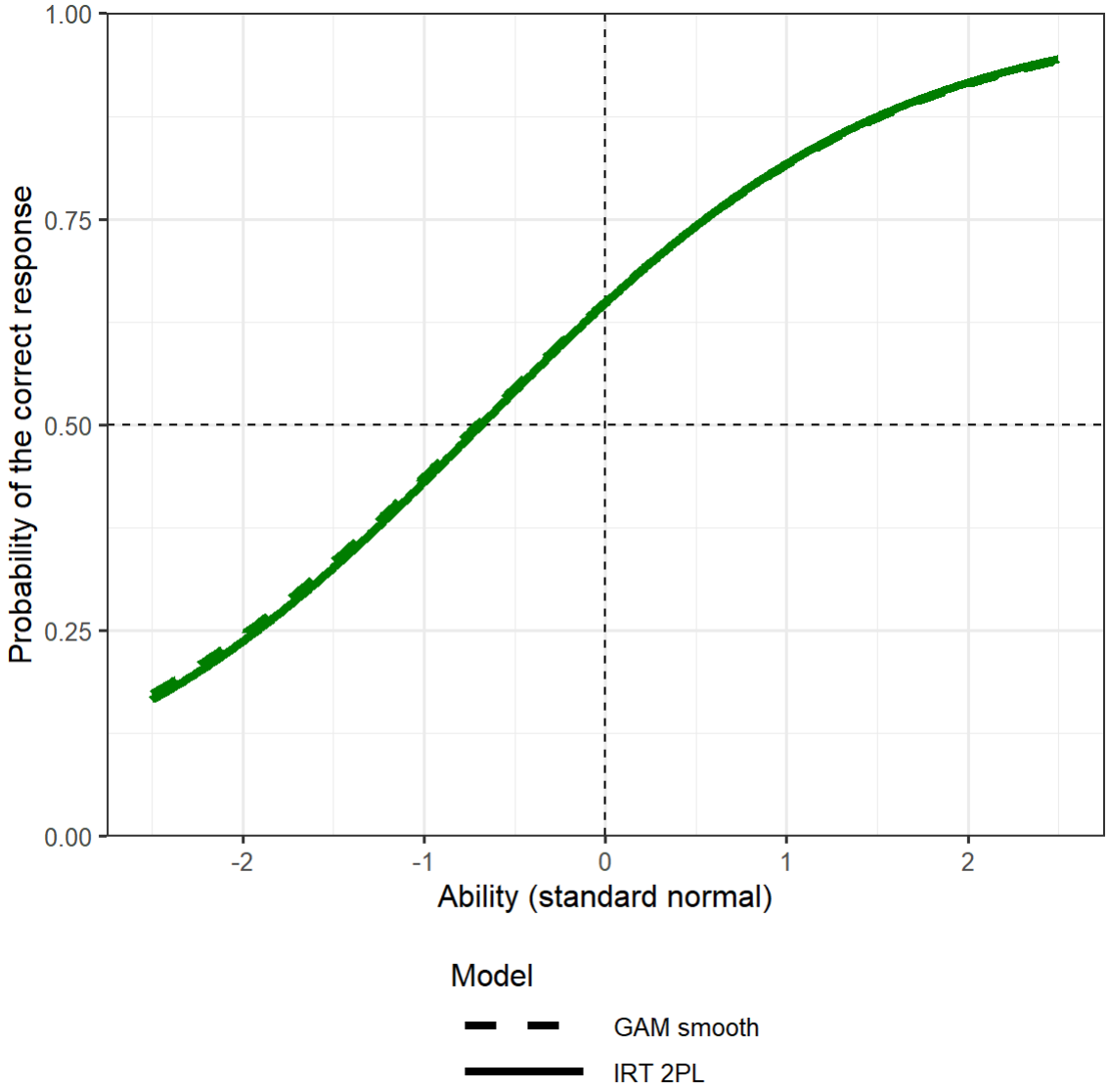
10.2.7.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.0	12.5	21.2	29.3	110.6	3.0	0.7
Middle quartiles	5.1	18.2	22.8	30.0	100.9	3.2	0.5
Upper quartile	9.4	15.6	20.3	27.8	83.7	3.1	0.5
Overall	4.0	16.3	21.9	29.3	110.6	3.1	0.5

10.2.8 SP71017 (TIMSS)

10.2.8.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	63.42%	299	TIMSS	MC	0.36	0.41	0.89	−0.68	0.005

10.2.8.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.75	0.253	1.000	0.008	N/A

10.2.8.3 Frequency of answers:

Answer	Frequency	Percent
Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A)	5	1.7%
Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	2	0.7%
Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B)	1	0.3%
Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	1	0.3%
Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	1	0.3%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A)	2	0.7%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B)	1	0.3%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	1	0.3%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	1	0.3%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A)	1	0.3%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B)	9	3.0%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	2	0.7%
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A)	1	0.3%

Answer	Frequency	Percent
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	1	0.3%
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	1	0.3%
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B)	2	0.7%
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	1	0.3%
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	3	1.0%
Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B)	3	1.0%
Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A)	3	1.0%
Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B) (correct)	189	63.4%
Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	3	1.0%
Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	64	21.5%
Overall	298	100.0%

10.2.8.4 Distractors functioning:

Answer	Frequency				Percent			
	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A)	3	2	0	5	6.7%	4.2%	0.0%	4.6%
Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	1	0	1	2	2.2%	0.0%	6.2%	1.8%
Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	1	0	0	1	2.2%	0.0%	0.0%	0.9%

Answer	Frequency				Percent			
	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	1	0	0	1	2.2%	0.0%	0.0%	0.9%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A)	2	0	0	2	4.4%	0.0%	0.0%	1.8%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B)	1	0	0	1	2.2%	0.0%	0.0%	0.9%

Answer	Frequency				Percent			
	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	1	0	0	1	2.2%	0.0%	0.0%	0.9%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A)	1	0	0	1	2.2%	0.0%	0.0%	0.9%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B)	5	2	2	9	11.1%	4.2%	12.5%	8.3%

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	2	0	0	2	4.4%	0.0%	0.0%	1.8%
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A)	1	0	0	1	2.2%	0.0%	0.0%	0.9%
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	1	0	0	1	2.2%	0.0%	0.0%	0.9%

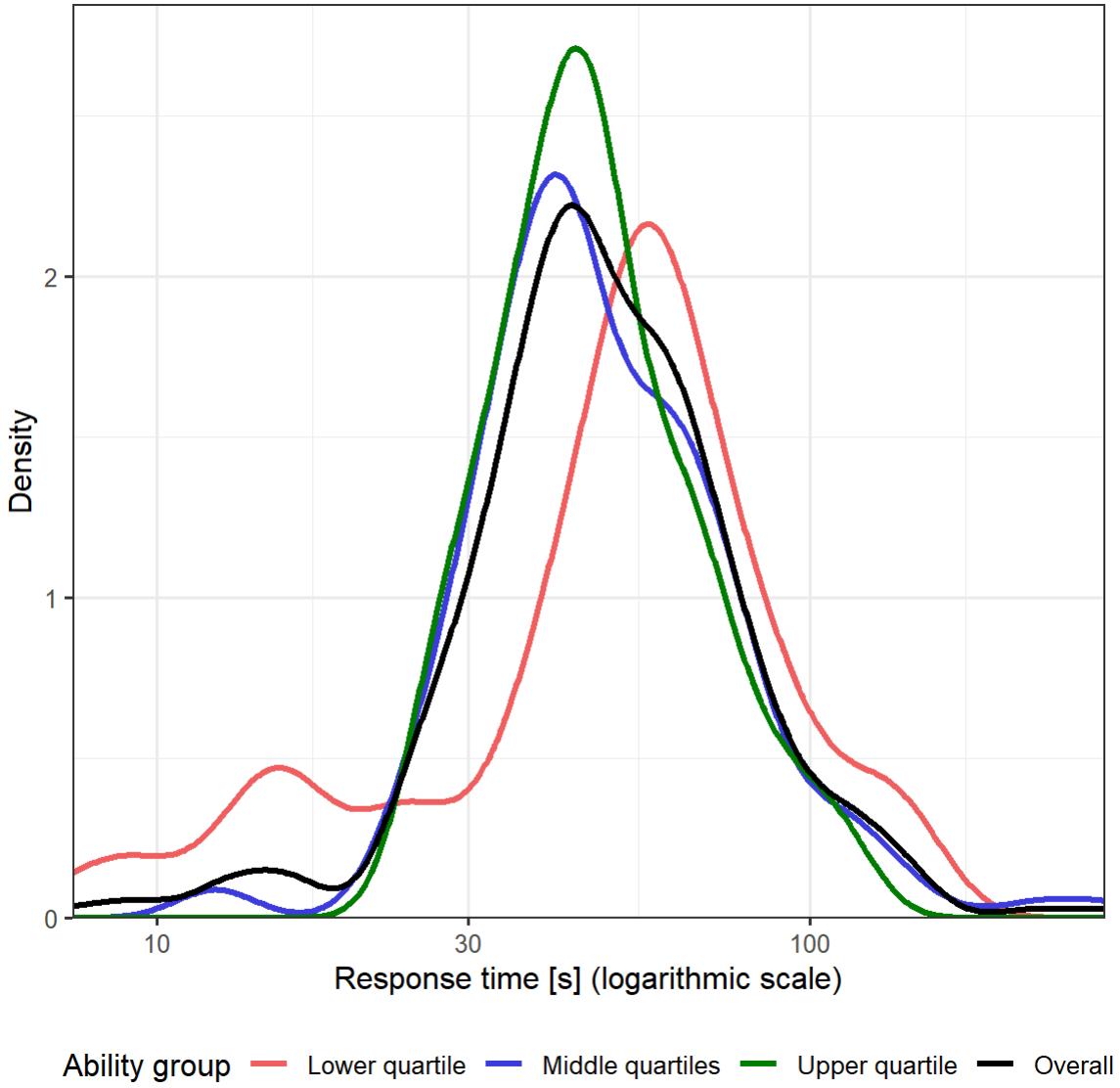
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	1	0	0	1	2.2%	0.0%	0.0%	0.9%
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	1	0	0	1	2.2%	0.0%	0.0%	0.9%
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	1	2	0	3	2.2%	4.2%	0.0%	2.8%

Answer	Frequency				Percent			
	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B)	1	2	0	3	2.2%	4.2%	0.0%	2.8%
Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A)	2	0	1	3	4.4%	0.0%	6.2%	2.8%
Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	2	1	0	3	4.4%	2.1%	0.0%	2.8%

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B), Tylko żywa kaczka (B)	17	36	11	64	37.8%	75.0%	68.8%	58.7%
Overall	45	48	16	109	100.0%	100.0%	100.0%	100.0%
Kaczka-zabawka i żywa kaczka (A), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B)	0	1	0	1	0.0%	2.1%	0.0%	0.9%
Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A)	0	1	0	1	0.0%	2.1%	0.0%	0.9%

Answer	Frequency				Percent			
	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B), Kaczka-zabawka i żywa kaczka (A), Tylko żywa kaczka (B)	0	1	1	2	0.0%	2.1%	6.2%	1.8%

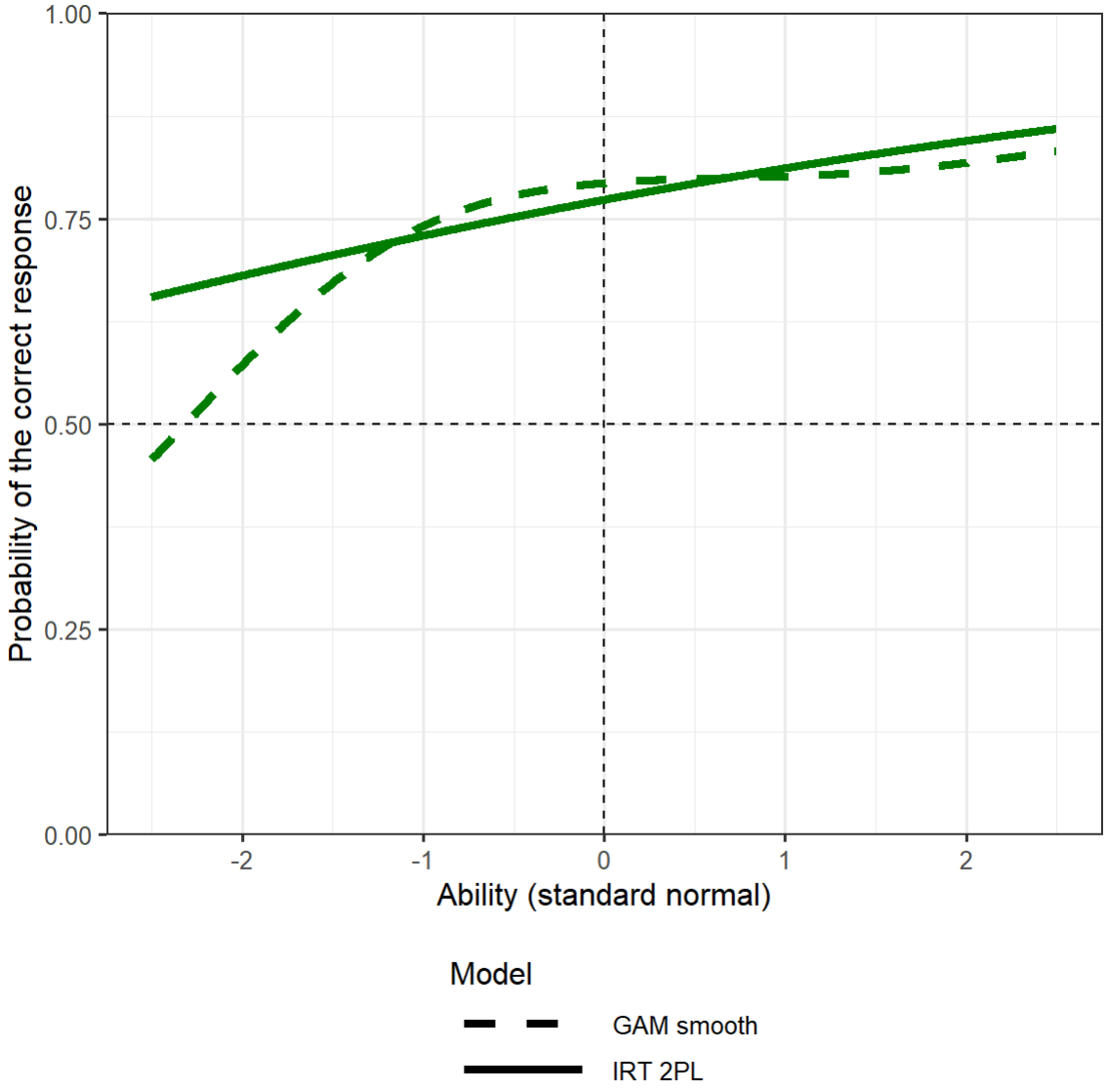
10.2.8.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	7.4	39.7	54.1	70.6	141.8	3.9	0.7
Middle quartiles	11.9	36.4	44.9	63.4	282.3	3.9	0.5
Upper quartile	26.2	36.5	45.0	58.8	114.3	3.8	0.4
Overall	7.4	36.5	46.7	63.4	282.3	3.9	0.5

10.2.9 S061071 (TIMSS)

10.2.9.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	77.18%	299	TIMSS	MC	0.11	0.14	0.23	−5.24	0.000

10.2.9.2 Gender DIF:

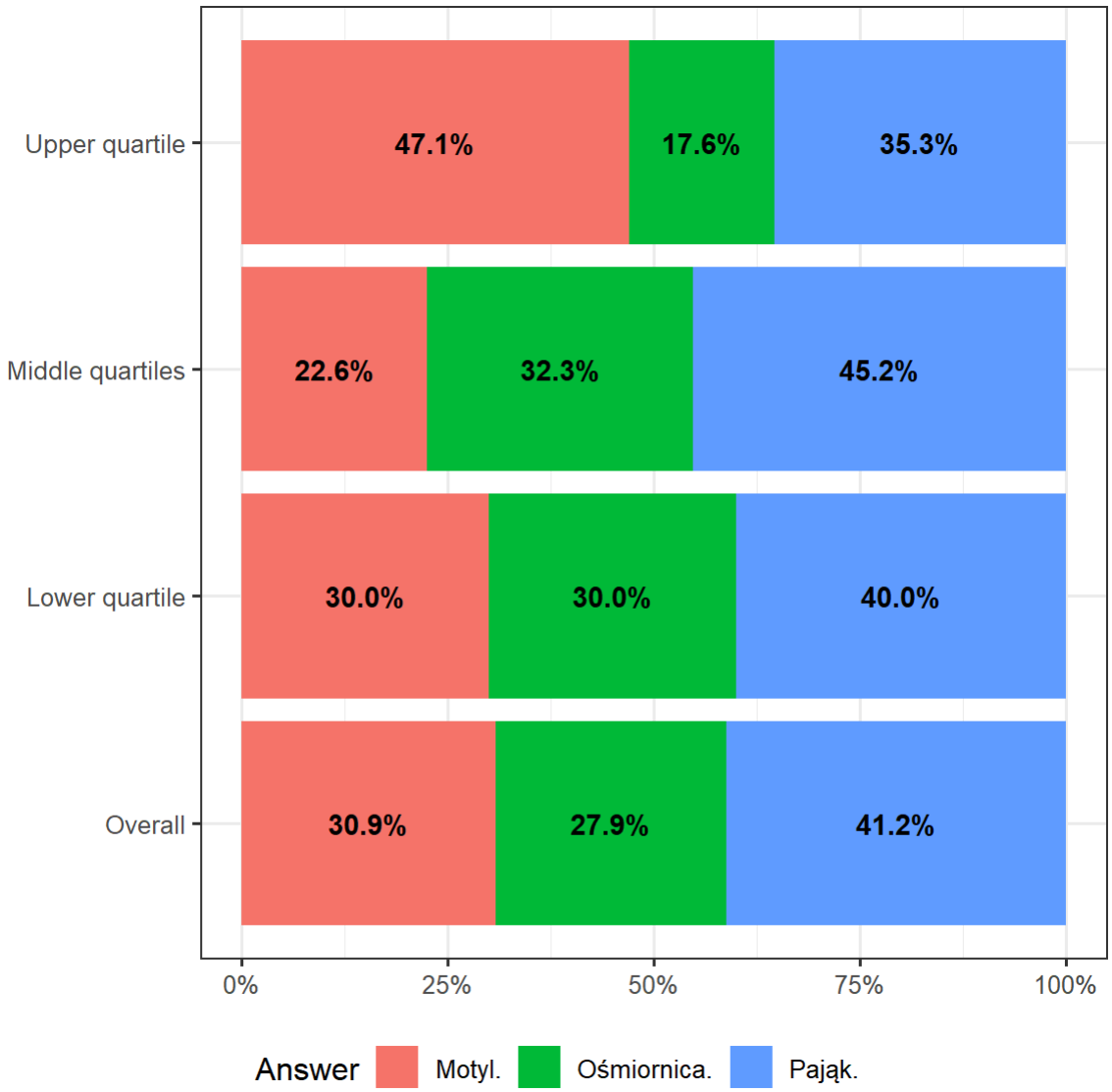
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
4.90	0.086	1.000	0.017	N/A

10.2.9.3 Frequency of answers:

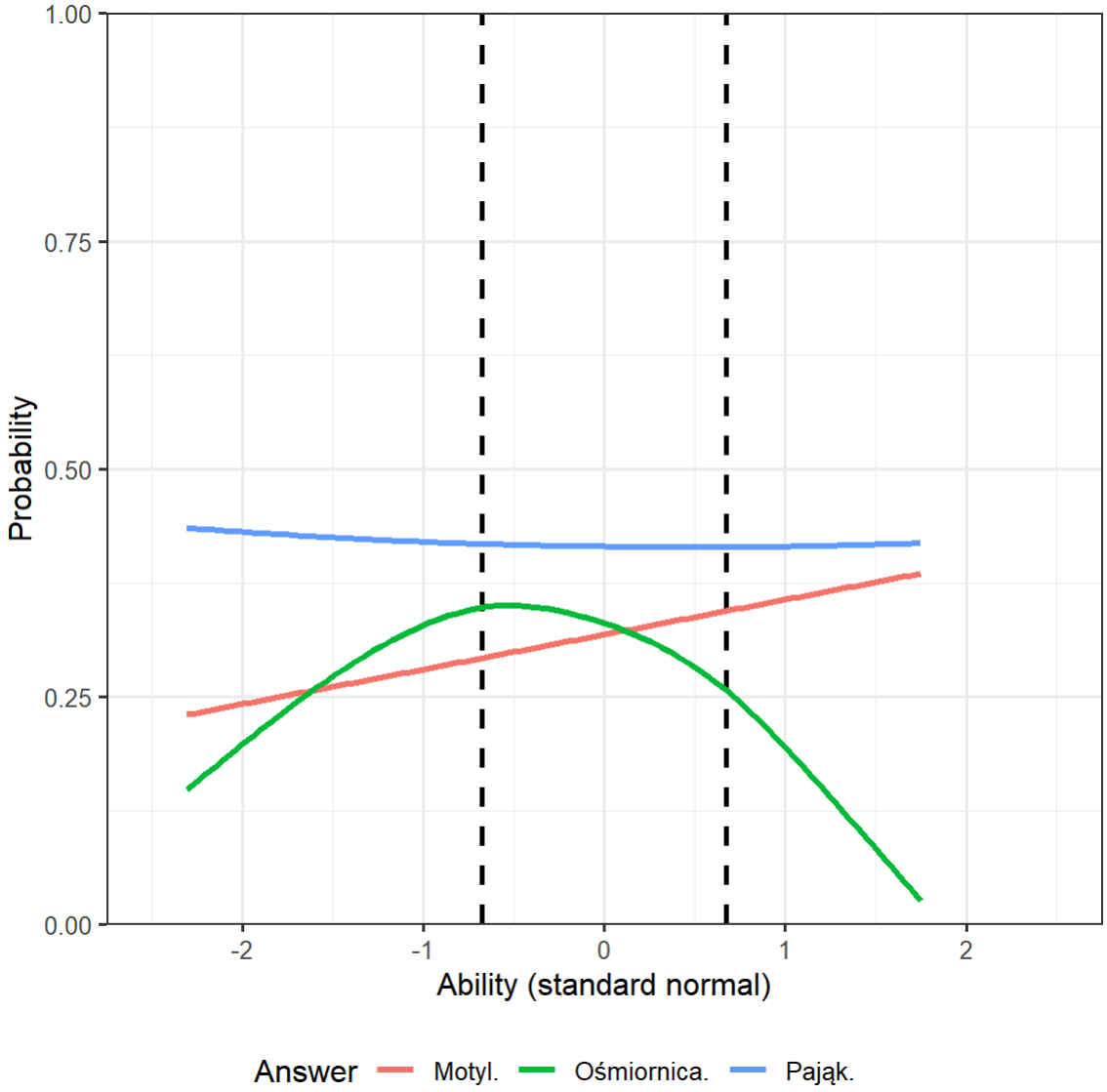
Answer	Frequency	Percent
Motyl.	21	7.0%
Ośmiornica.	19	6.4%
Pająk.	28	9.4%
Żaba. (correct)	230	77.2%
Overall	298	100.0%

10.2.9.4 Distractors functioning:

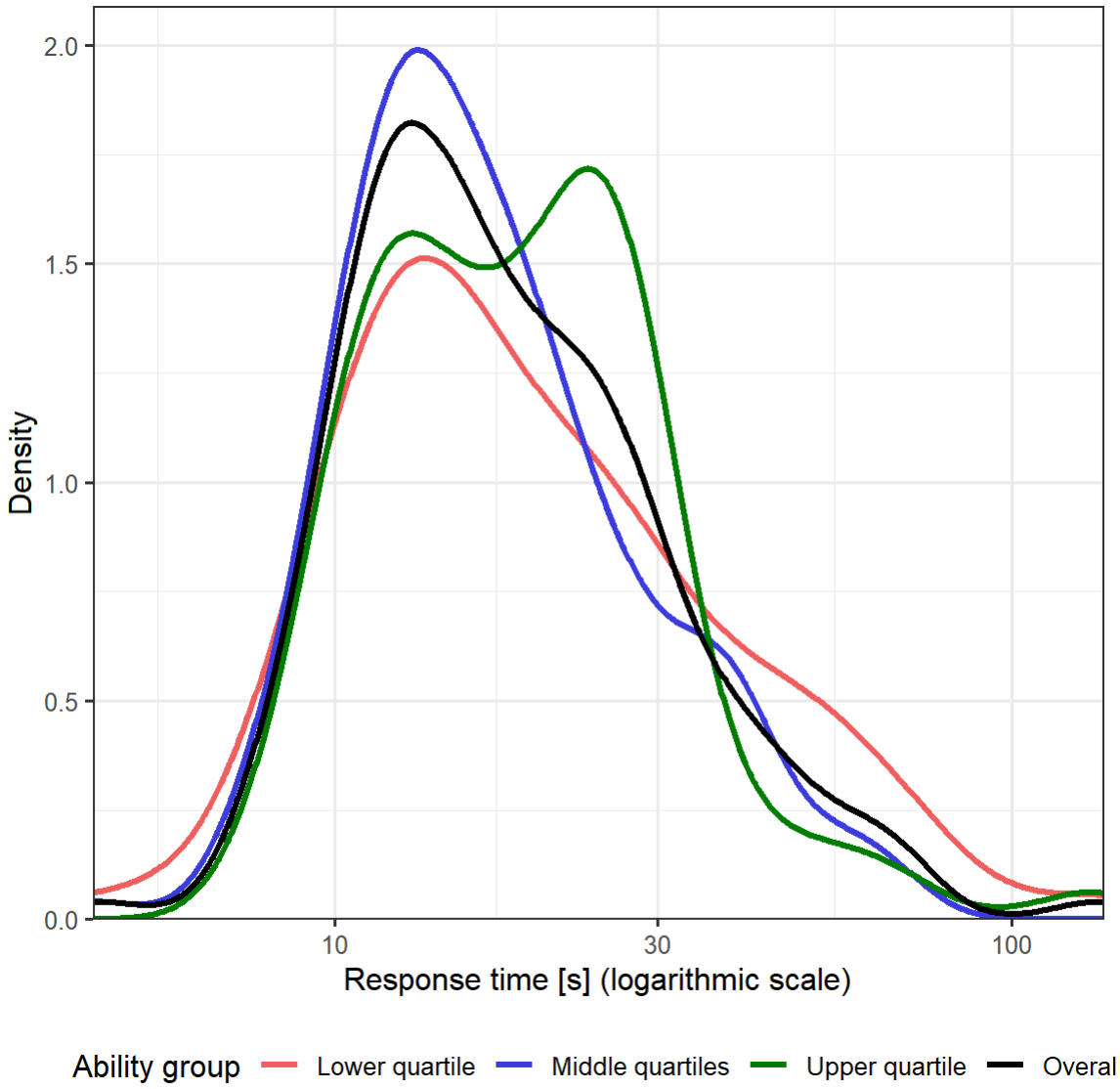
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Motyl.	6	7	8	21	30.0%	22.6%	47.1%	30.9%
Ośmiornica.	6	10	3	19	30.0%	32.3%	17.6%	27.9%
Pająk.	8	14	6	28	40.0%	45.2%	35.3%	41.2%
Overall	20	31	17	68	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



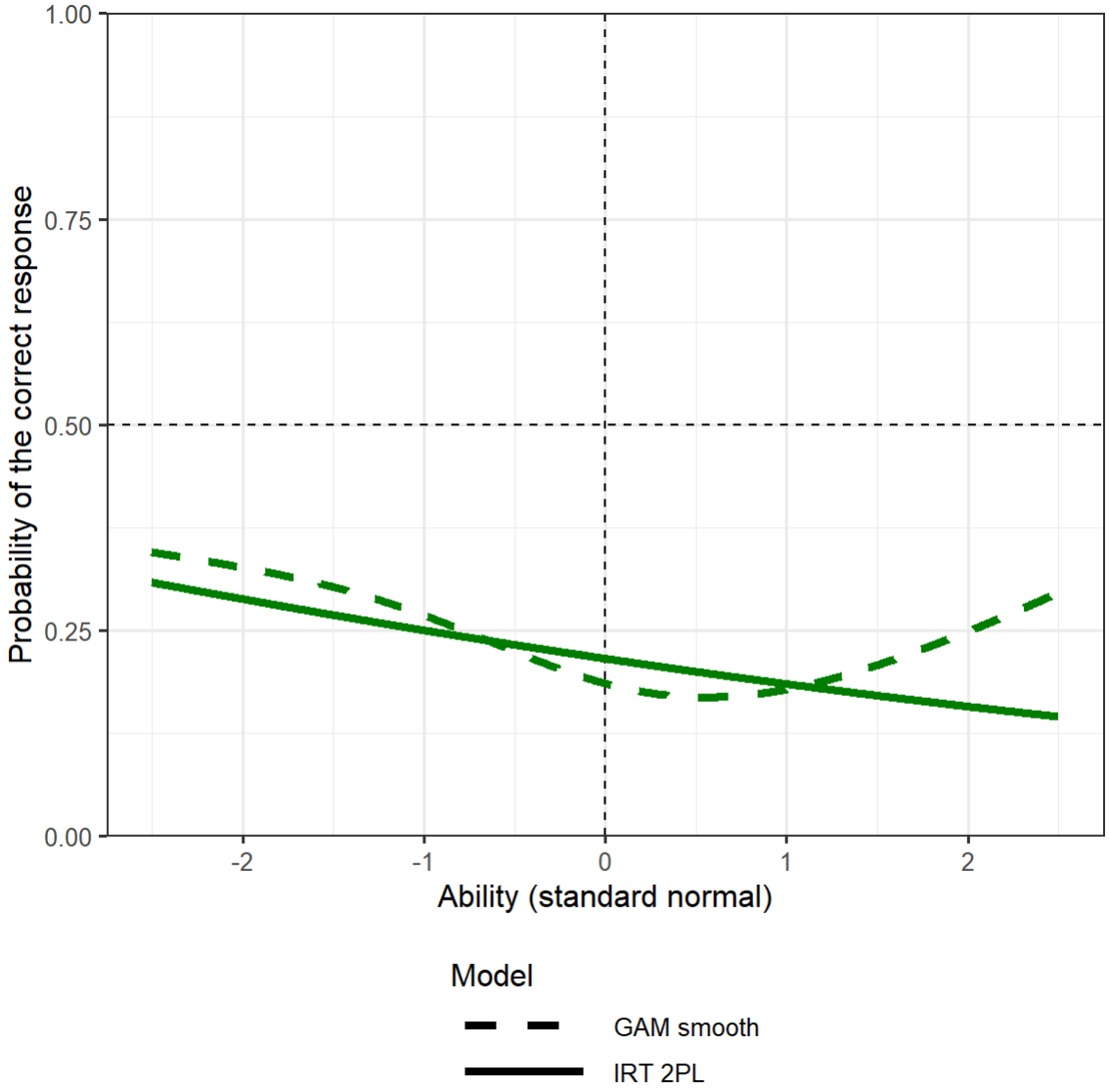
10.2.9.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.6	12.0	16.6	28.4	136.6	3.0	0.6
Middle quartiles	4.4	12.0	15.5	22.7	66.4	2.8	0.5
Upper quartile	8.5	12.6	19.7	25.9	131.2	2.9	0.5
Overall	4.4	12.0	16.6	25.1	136.6	2.9	0.5

10.2.10 S090058 (LLM)

10.2.10.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS and generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSI
1	21.79%	281	LLM teachers' selection	MC	−0.08	−0.10	−0.19	−6.68	0.0

10.2.10.2 Gender DIF:

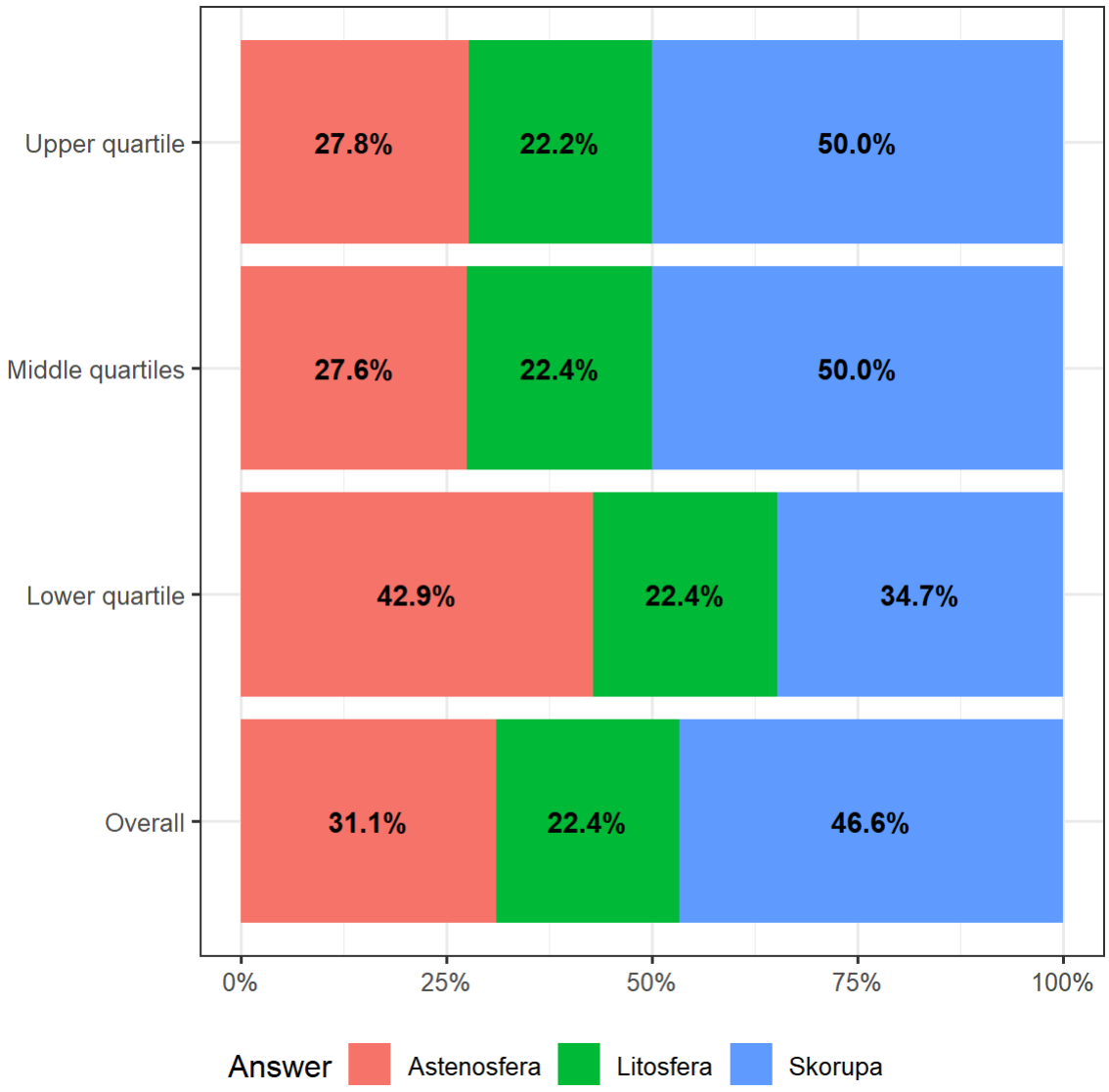
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.33	0.515	1.000	0.005	N/A

10.2.10.3 Frequency of answers:

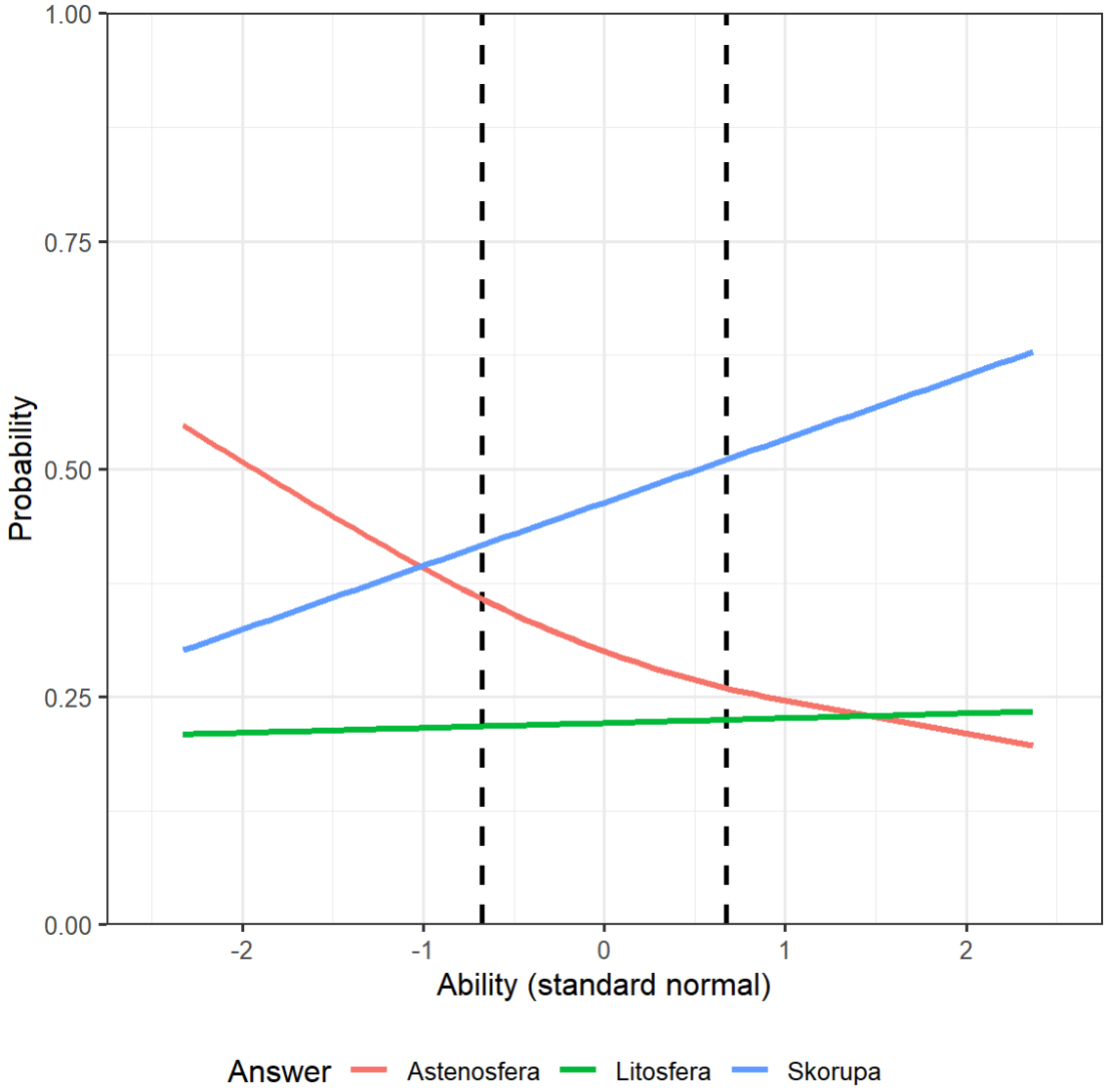
Answer	Frequency	Percent
Astenosfera	68	24.3%
Litosfera	49	17.5%
Płaszcz (correct)	61	21.8%
Skorupa	102	36.4%
Overall	280	100.0%

10.2.10.4 Distractors functioning:

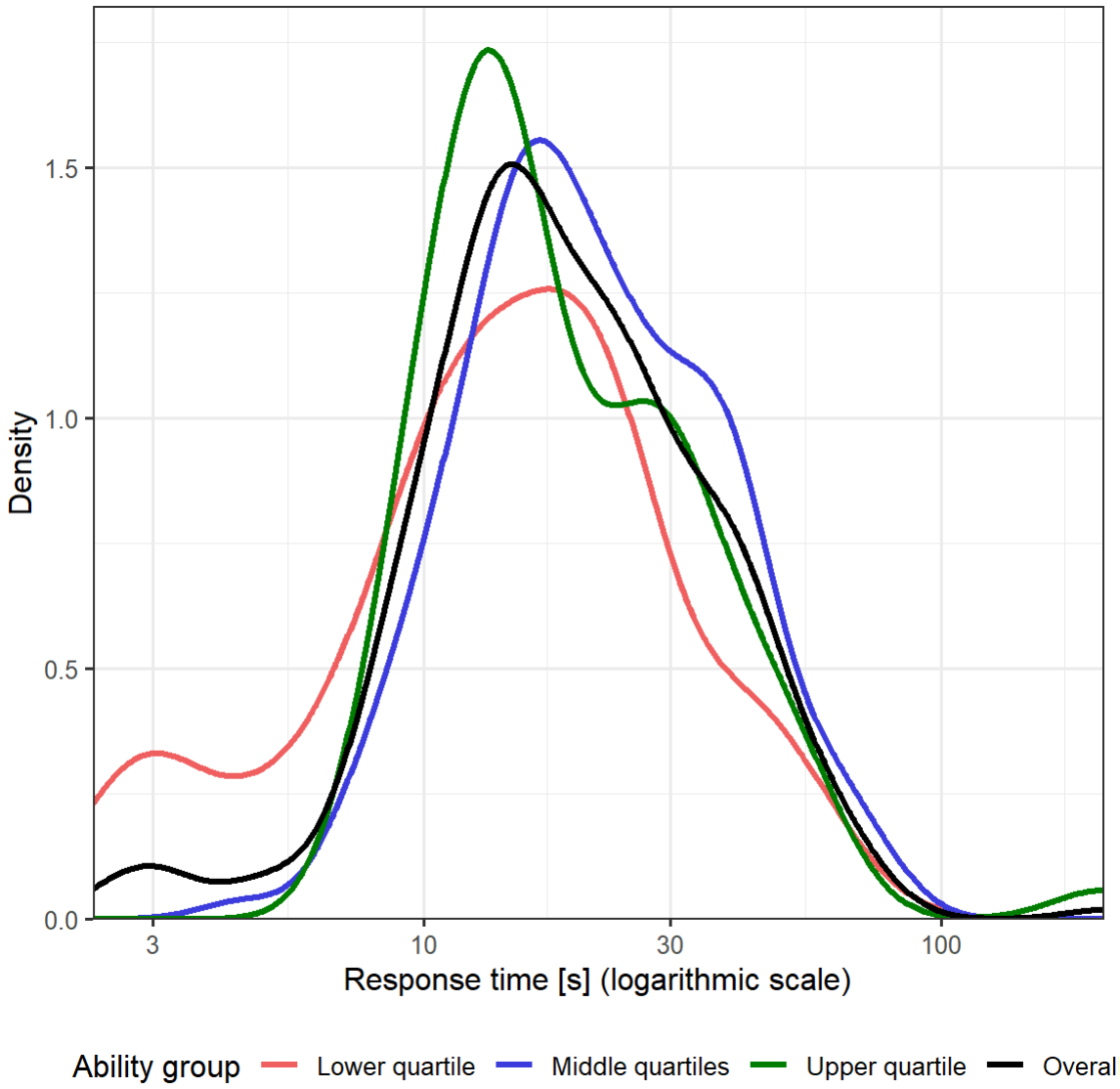
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Astenosfera	21	32	15	68	42.9%	27.6%	27.8%	31.1%
Litosfera	11	26	12	49	22.4%	22.4%	22.2%	22.4%
Skorupa	17	58	27	102	34.7%	50.0%	50.0%	46.6%
Overall	49	116	54	219	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



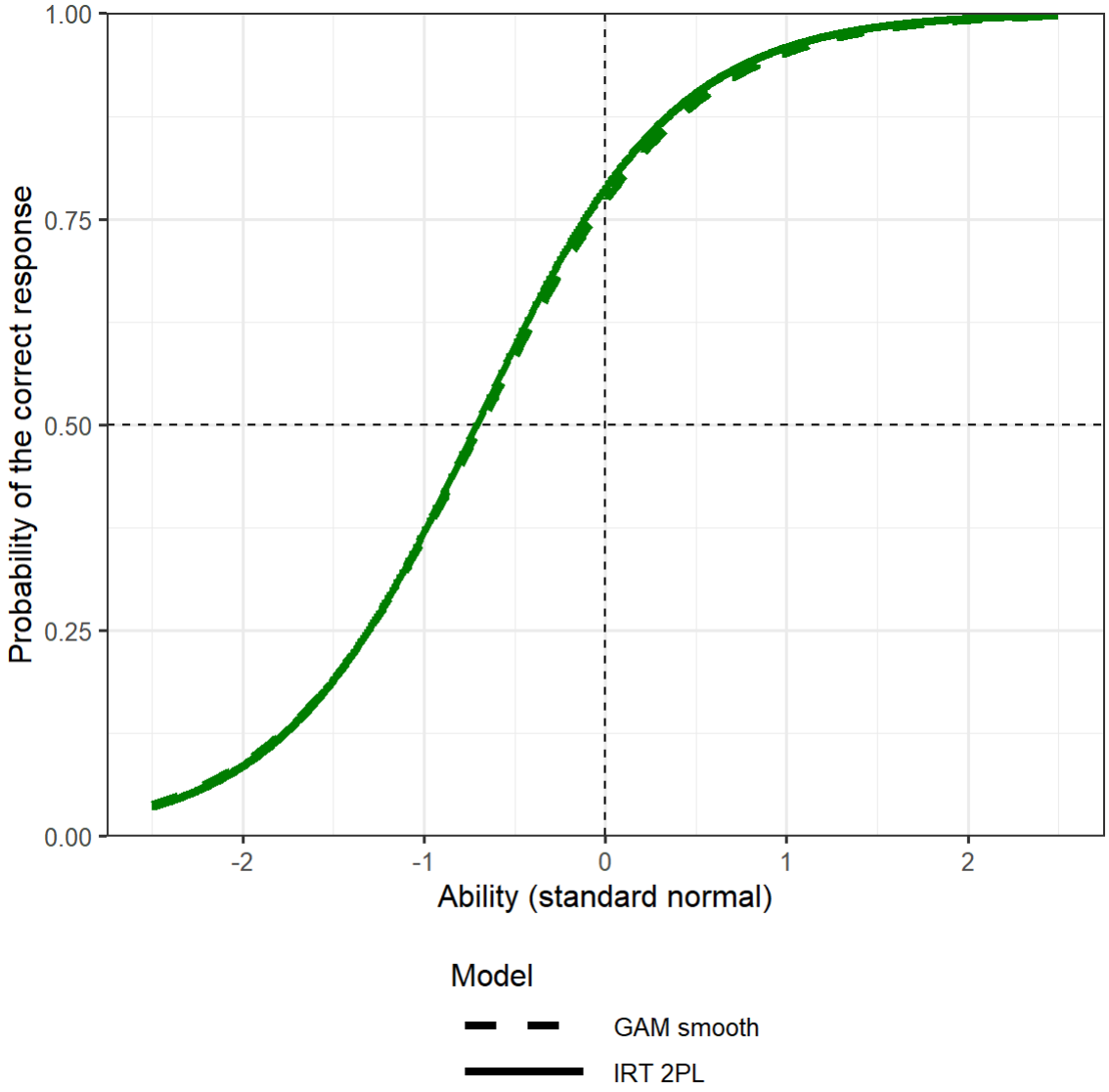
10.2.10.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.3	9.2	15.1	23.3	65.2	2.6	0.8
Middle quartiles	4.4	14.6	20.1	32.6	79.3	3.0	0.6
Upper quartile	8.3	12.8	16.0	28.5	205.8	2.9	0.6
Overall	2.3	12.5	17.8	28.6	205.8	2.9	0.6

10.2.11 SP71147 (LLM)

10.2.11.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	69.08%	305	TIMSS	MC	0.52	0.54	1.84	−0.71	0.025

10.2.11.2 Gender DIF:

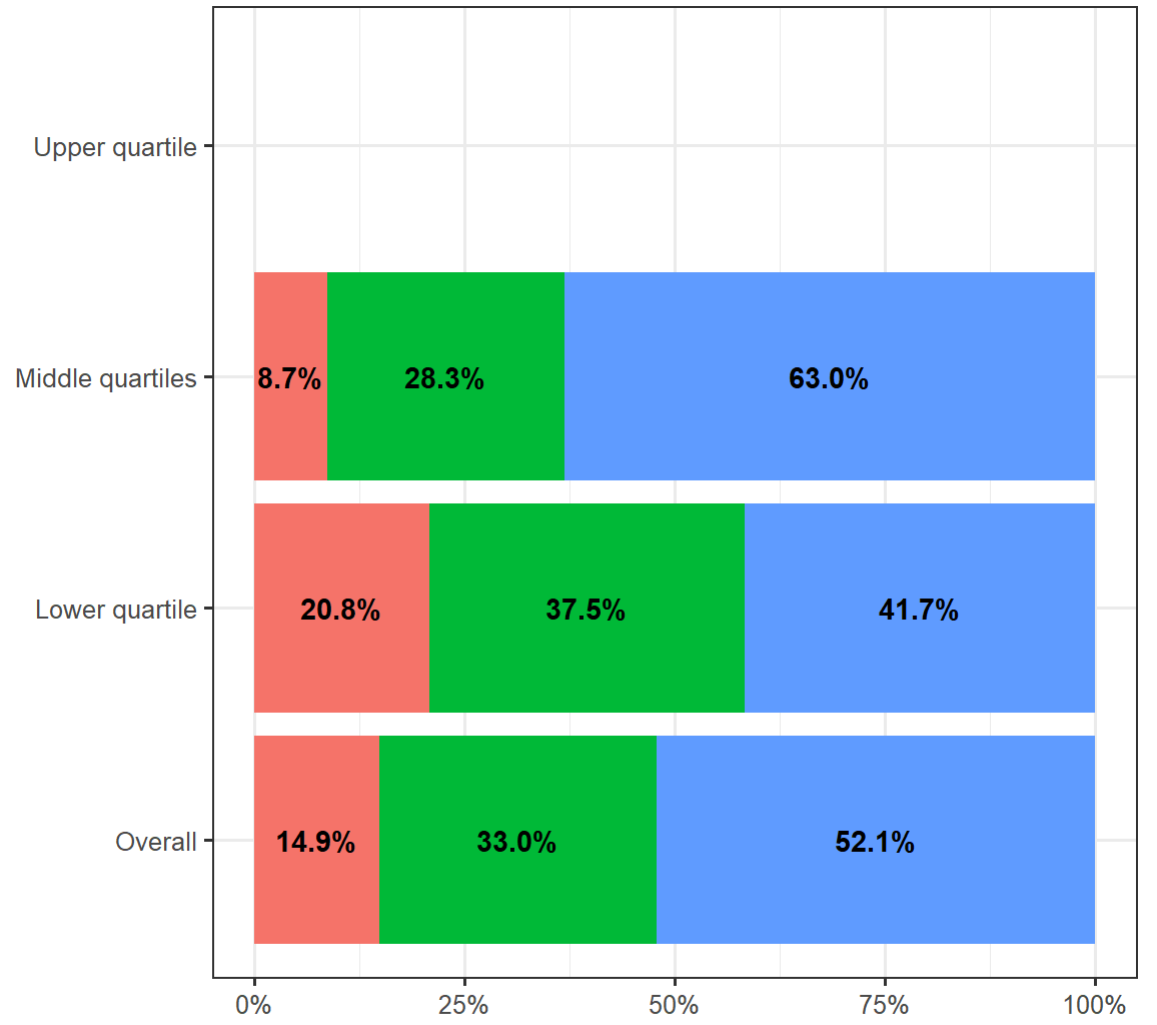
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
7.14	0.028	1.000	0.021	N/A

10.2.11.3 Frequency of answers:

Answer	Frequency	Percent
Grawitacja działająca na skrzynię Teresy jest dużo silniejsza.	14	4.6%
Kółka wózka zmniejszają siłę potrzebną do przeniesienia skrzyni Marysi. (correct)	210	69.1%
Opór powietrza działający na skrzynię Teresy jest dużo większy.	31	10.2%
Wózek zwiększa siłę magnetyczną działającą na skrzynię Marysi.	49	16.1%
Overall	304	100.0%

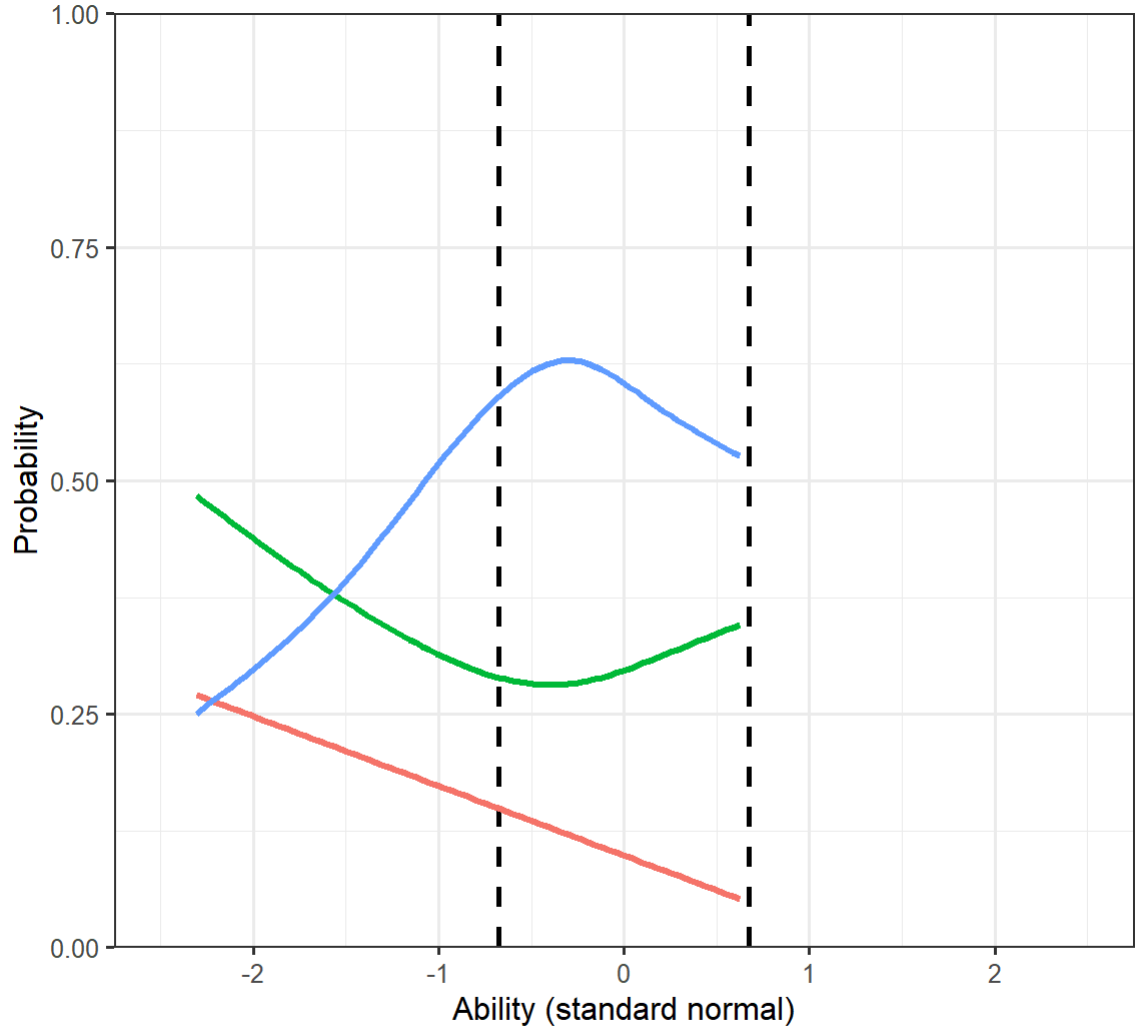
10.2.11.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Grawitacja działająca na skrzynię Teresy jest dużo silniejsza.	10	4	0	14	20.8%	8.7%	0.0%	14.9%
Opór powietrza działający na skrzynię Teresy jest dużo większy.	18	13	0	31	37.5%	28.3%	0.0%	33.0%
Wózek zwiększa siłę magnetyczną działającą na skrzynię Marysi.	20	29	0	49	41.7%	63.0%	0.0%	52.1%
Overall	48	46	0	94	100.0%	100.0%	0.0%	100.0%



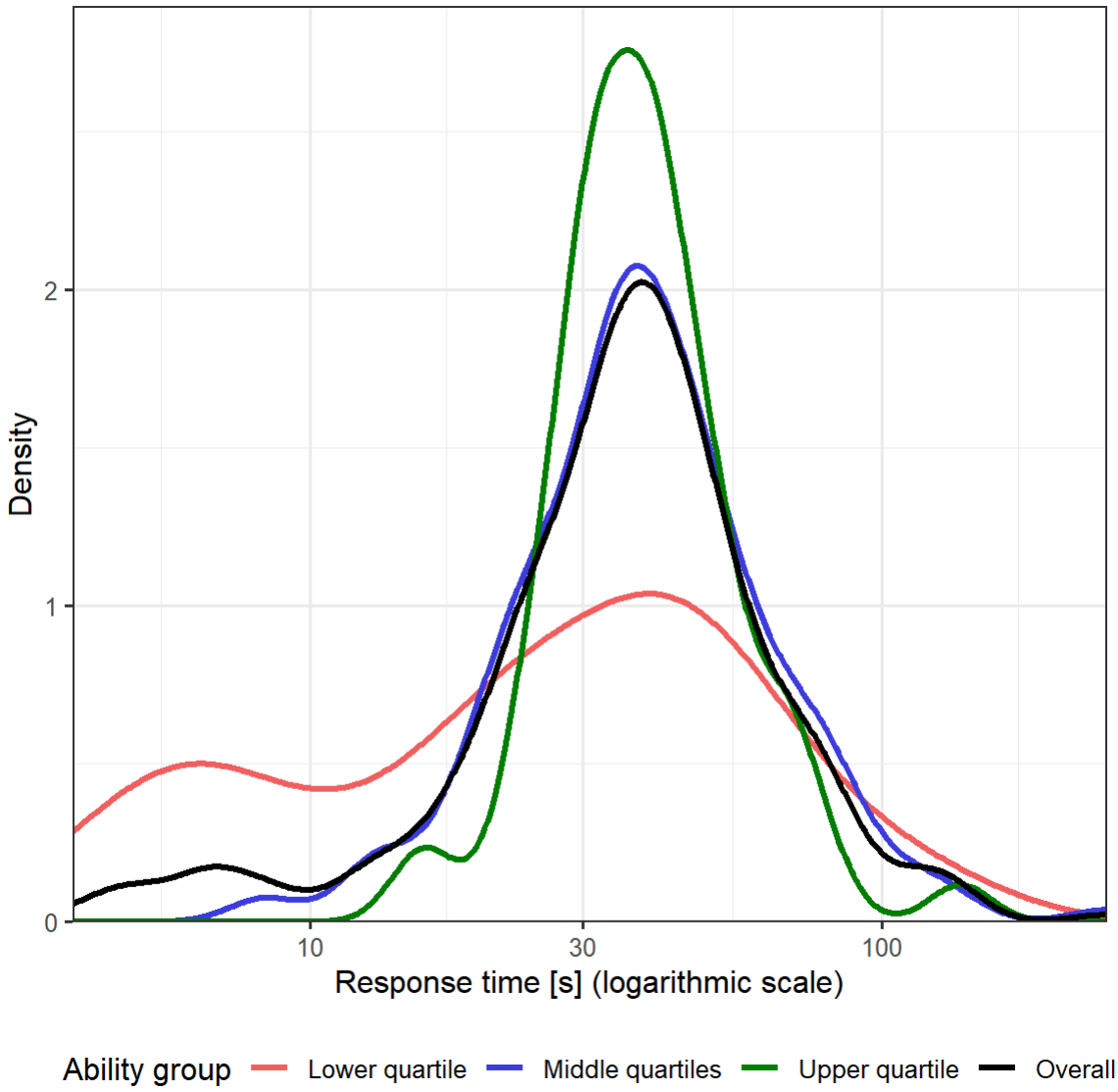
st dużo silniejsza. Opór powietrza działający na skrzynię Teresy jest dużo większy. Wóze

```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



st dużo silniejsza. Opór powietrza działający na skrzynię Teresy jest dużo większy. Wóze

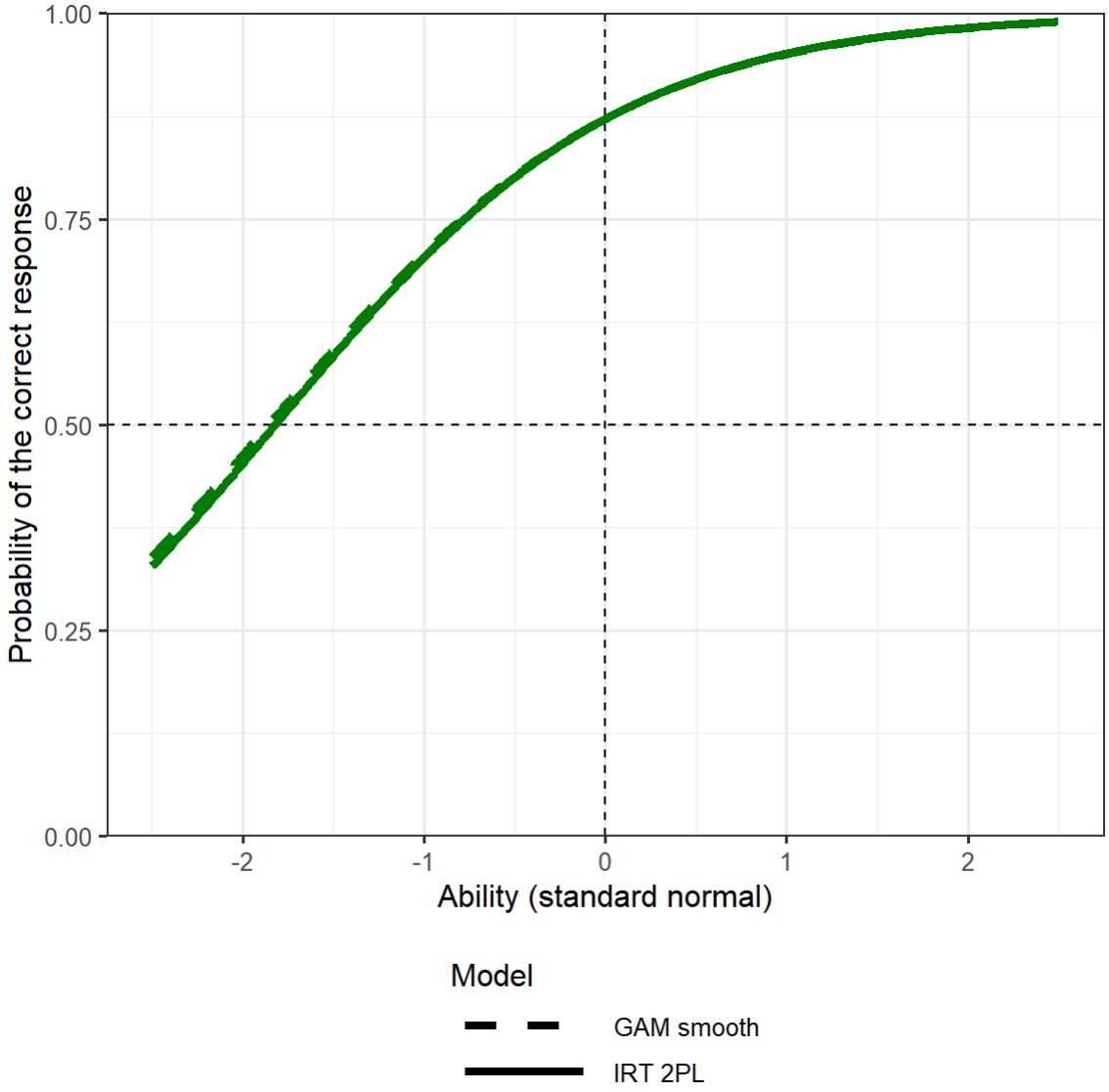
10.2.11.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.8	12.7	27.7	46.4	128.7	3.2	0.9
Middle quartiles	8.2	28.2	37.8	50.8	246.4	3.6	0.5
Upper quartile	15.4	30.9	38.7	47.6	137.2	3.6	0.4
Overall	3.8	25.8	37.0	48.9	246.4	3.5	0.6

10.2.12 S890038 (LLM)

10.2.12.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	82.86%	281	LLM complex setup	MC	0.35	0.19	1.05	−1.82	0.00

10.2.12.2 Gender DIF:

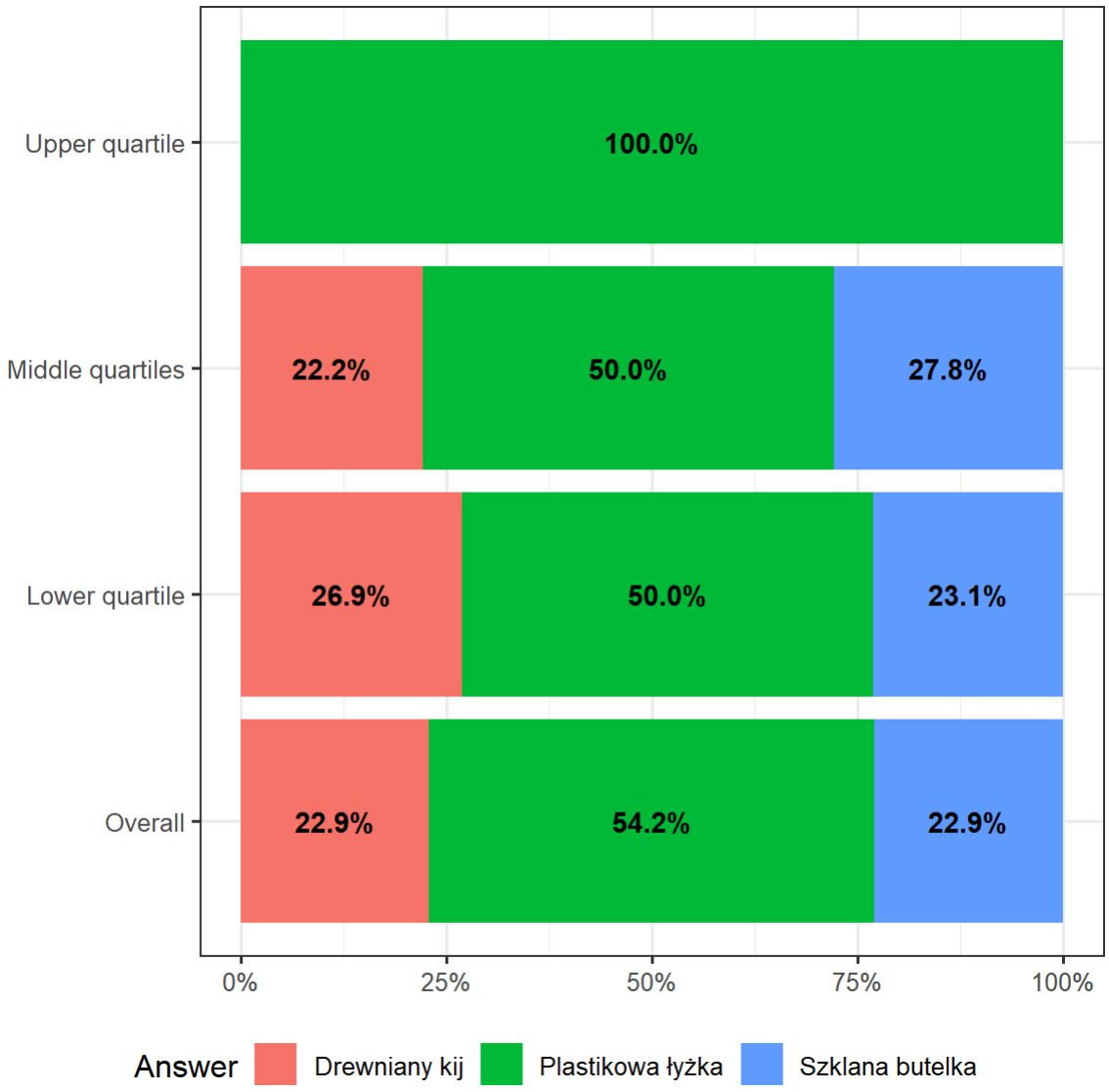
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
11.19	0.004	0.330	0.048	N/A

10.2.12.3 Frequency of answers:

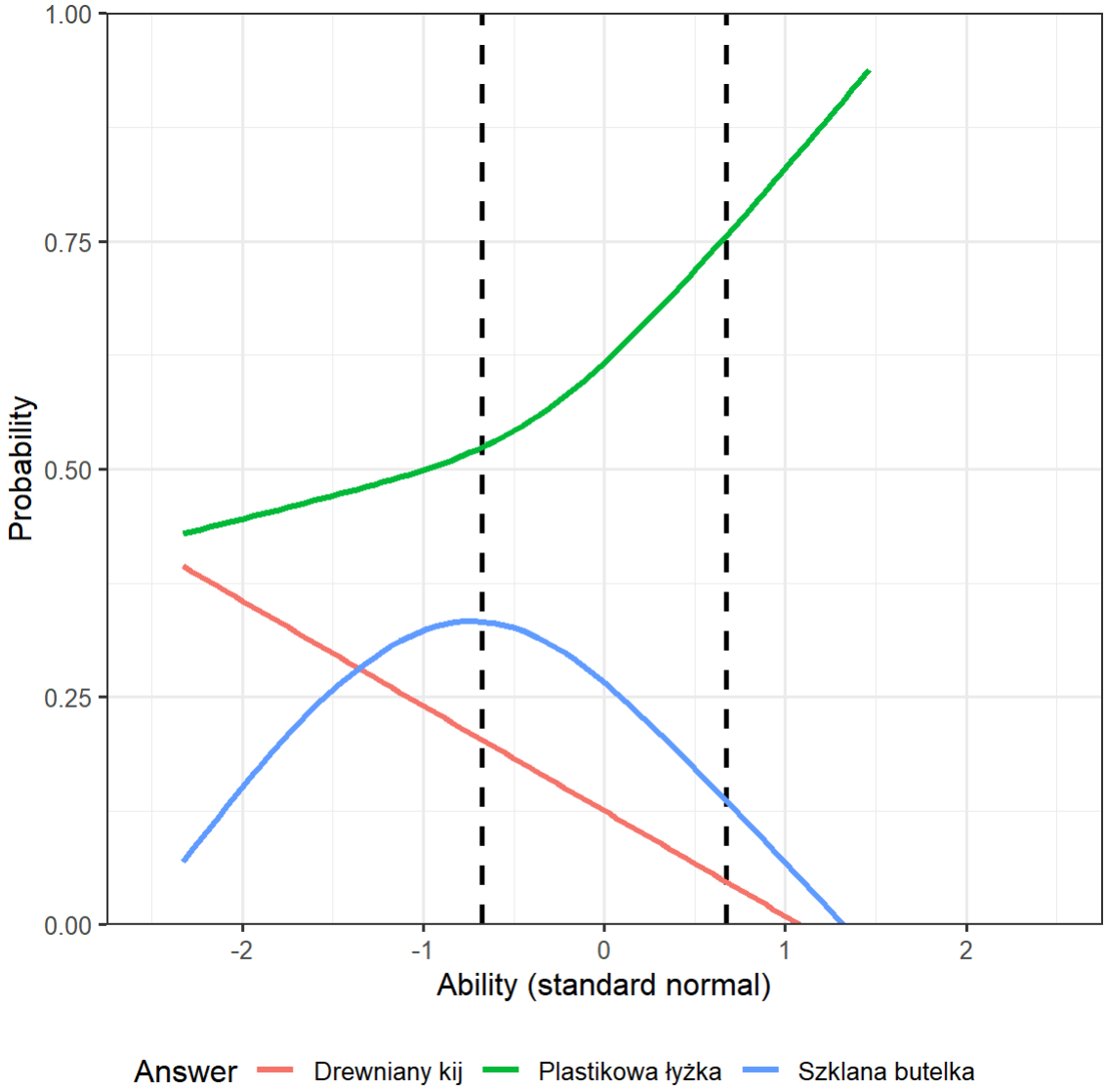
Answer	Frequency	Percent
Drewniany kij	11	3.9%
Metalowy gwóźdź (correct)	232	82.9%
Plastikowa łyżka	26	9.3%
Szklana butelka	11	3.9%
Overall	280	100.0%

10.2.12.4 Distractors functioning:

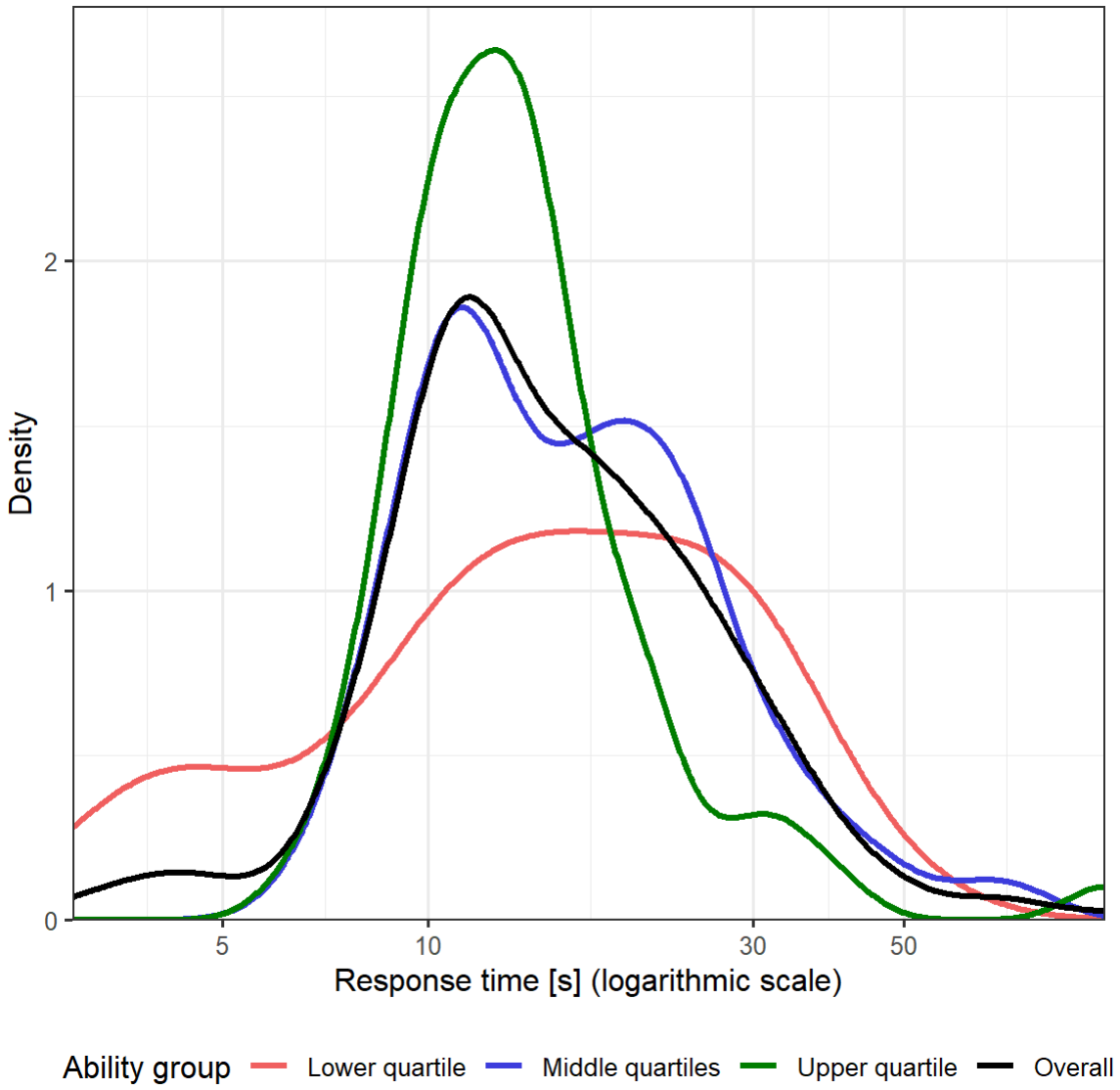
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Drewniany kij	7	4	0	11	26.9%	22.2%	0.0%	22.9%
Plastikowa łyżka	13	9	4	26	50.0%	50.0%	100.0%	54.2%
Szklana butelka	6	5	0	11	23.1%	27.8%	0.0%	22.9%
Overall	26	18	4	48	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



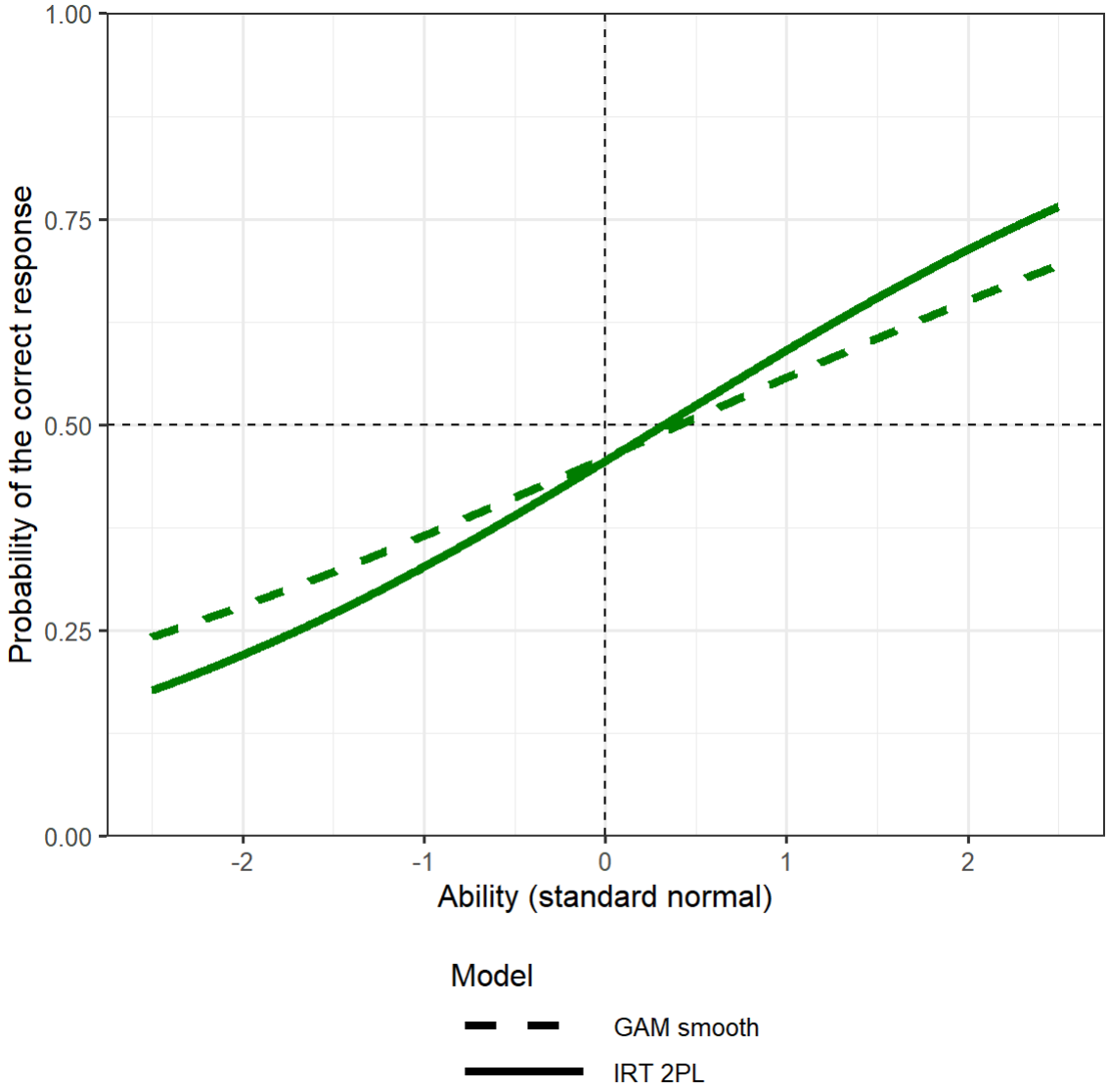
10.2.12.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	9.5	15.5	24.9	52.3	2.6	0.7
Middle quartiles	6.5	11.1	15.8	22.6	72.1	2.8	0.5
Upper quartile	6.4	10.3	12.8	16.2	98.4	2.6	0.4
Overall	3.0	10.6	14.1	21.3	98.4	2.7	0.5

10.2.13 SP71268 (LLM)

10.2.13.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	46.31%	299	TIMSS	MC	0.18	0.20	0.54	0.33	0.000

10.2.13.2 Gender DIF:

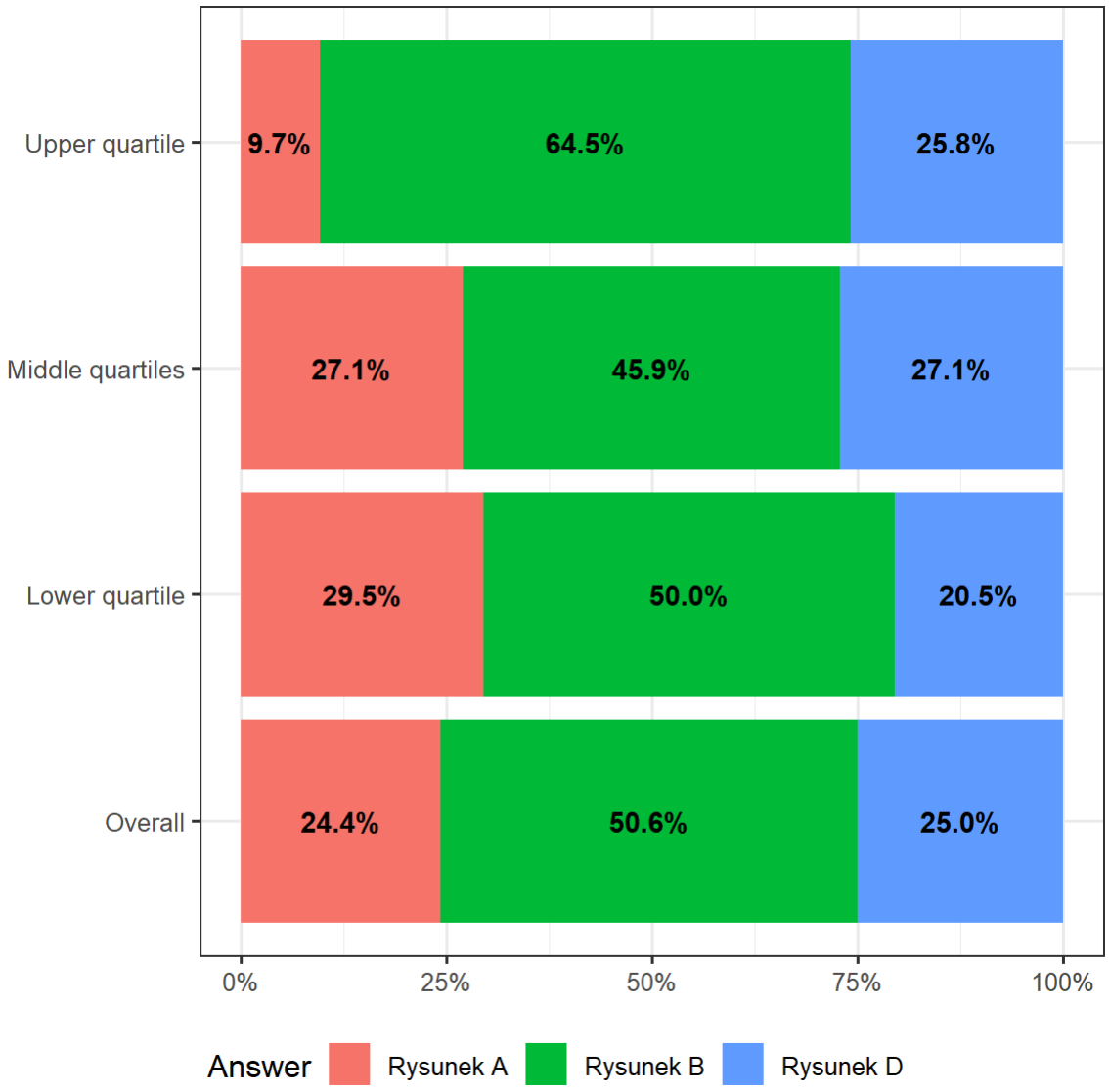
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
6.83	0.033	1.000	0.019	N/A

10.2.13.3 Frequency of answers:

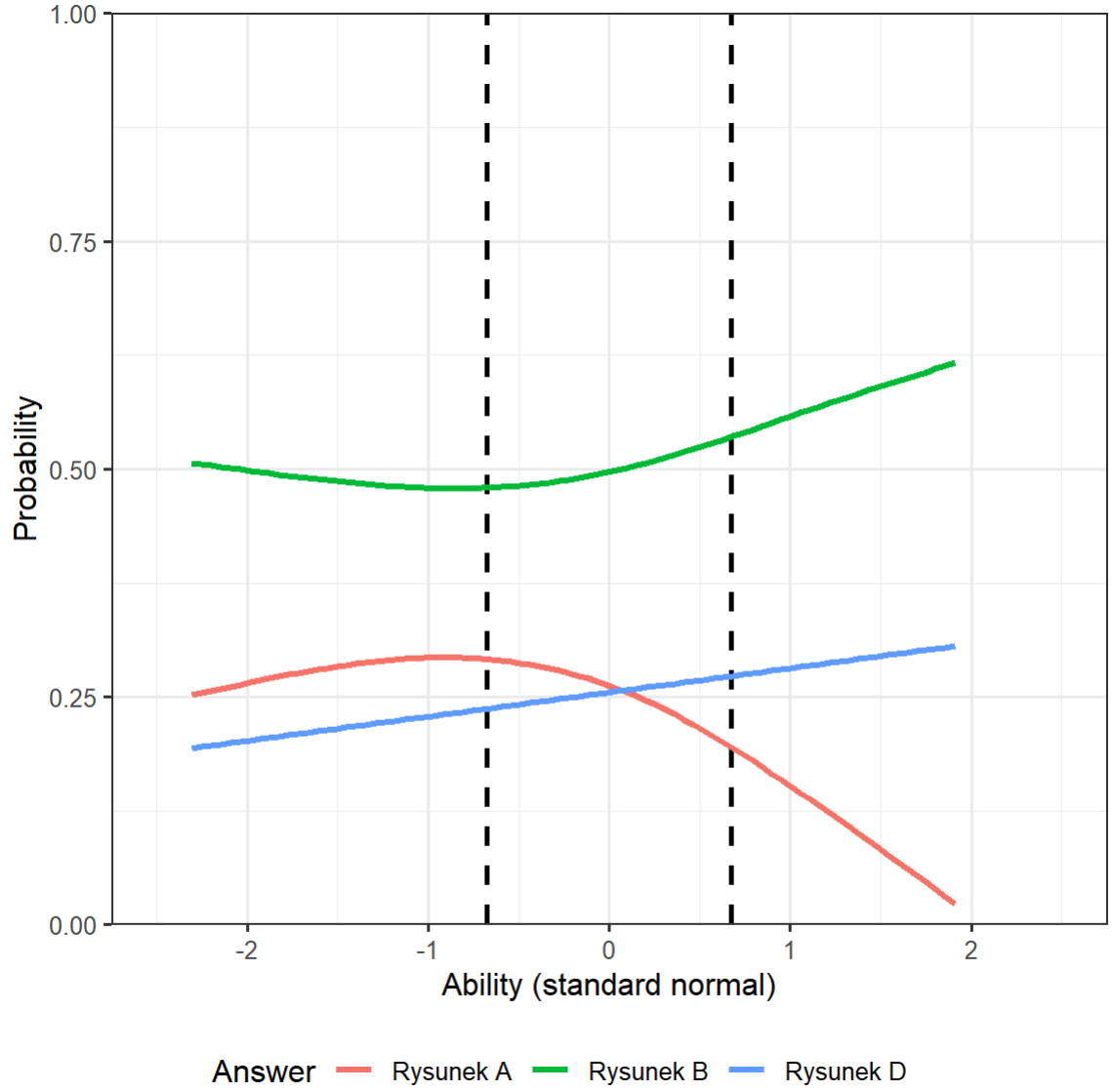
Answer	Frequency	Percent
Rysunek A	39	13.1%
Rysunek B	81	27.2%
Rysunek C (correct)	138	46.3%
Rysunek D	40	13.4%
Overall	298	100.0%

10.2.13.4 Distractors functioning:

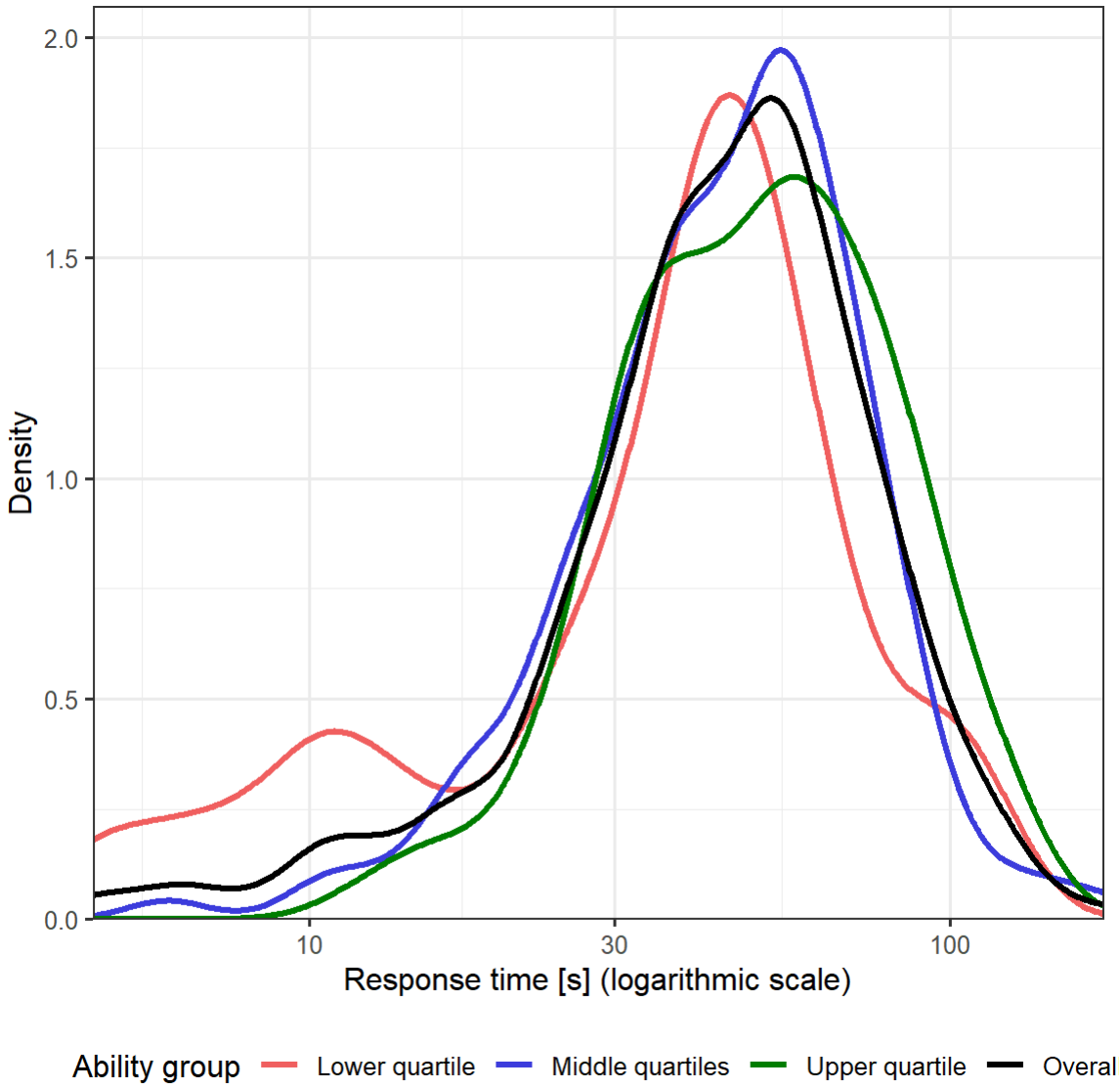
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Rysunek A	13	23	3	39	29.5%	27.1%	9.7%	24.4%
Rysunek B	22	39	20	81	50.0%	45.9%	64.5%	50.6%
Rysunek D	9	23	8	40	20.5%	27.1%	25.8%	25.0%
Overall	44	85	31	160	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



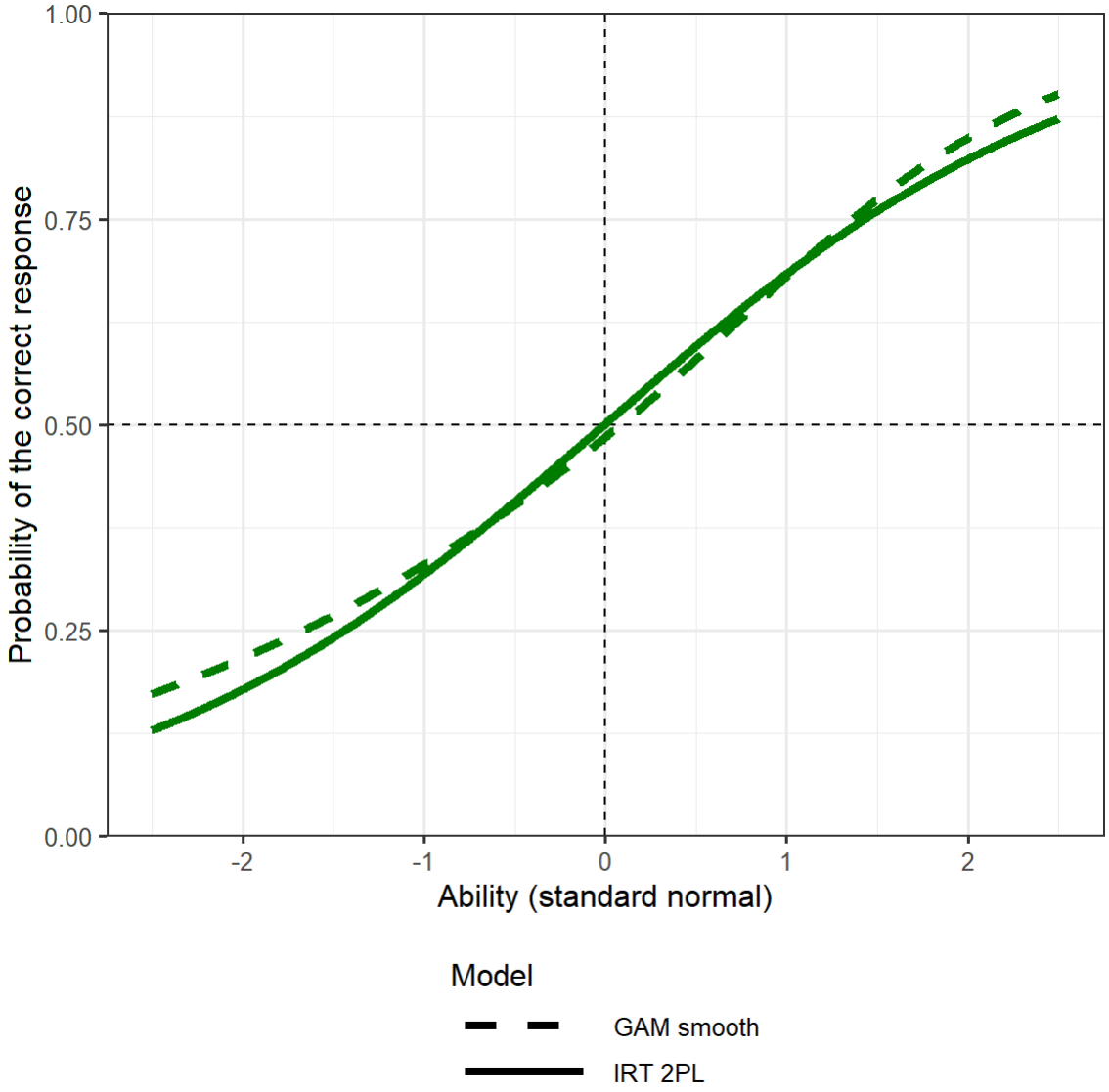
10.2.13.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.6	27.0	41.5	52.5	112.6	3.5	0.8
Middle quartiles	6.0	33.5	47.4	60.9	173.8	3.8	0.5
Upper quartile	12.6	34.9	51.4	69.9	127.9	3.9	0.5
Overall	4.6	32.4	46.8	61.6	173.8	3.7	0.6

10.2.14 S051188 (LLM)

10.2.14.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	51.02%	294	TIMSS	MC	0.32	0.26	0.77	0.00	0.009

10.2.14.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.69	0.261	1.000	0.008	N/A

10.2.14.3 Frequency of answers:

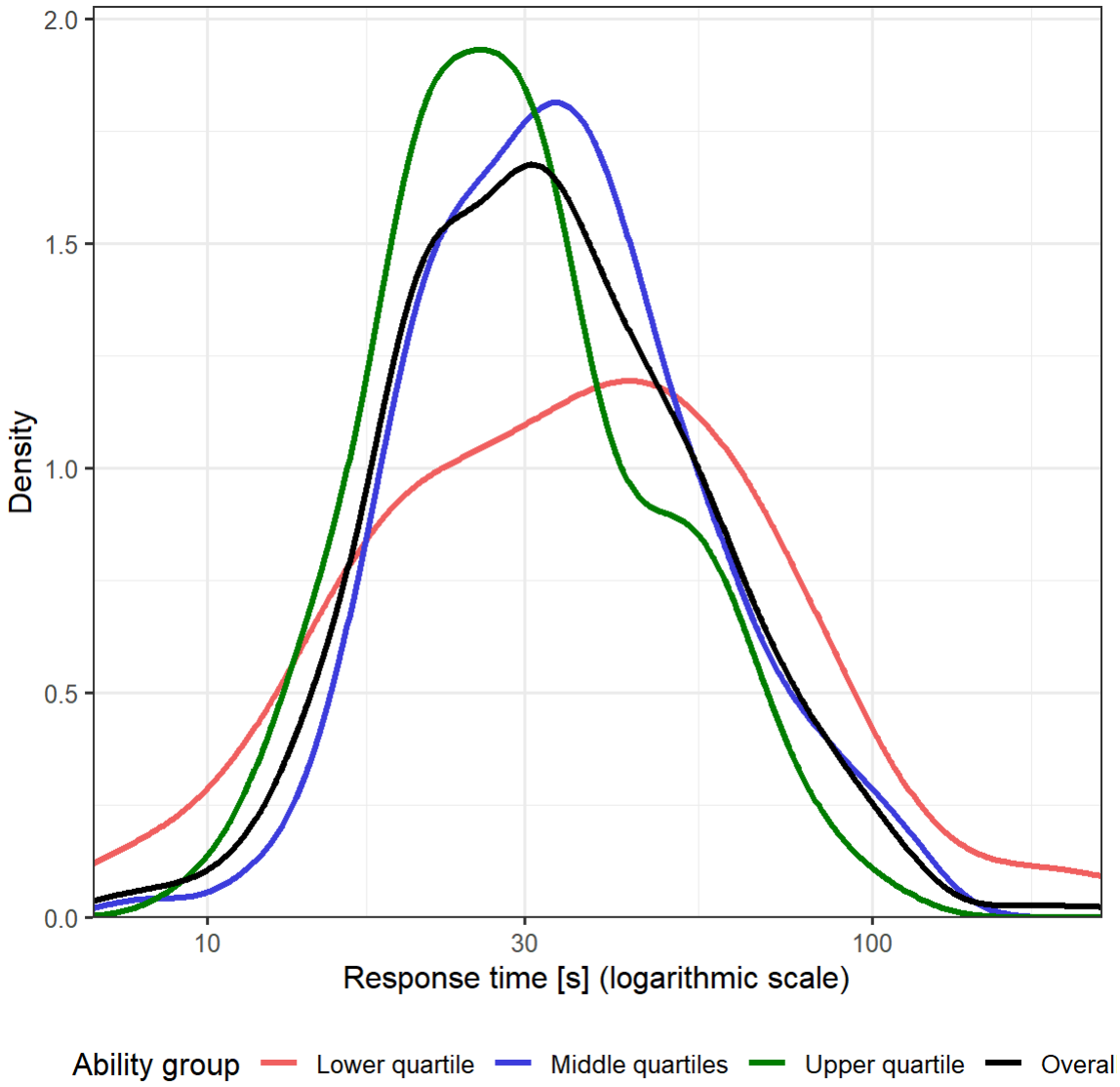
Answer	Frequency	Percent
Nie, Nie, Nie, Nie, Nie	2	0.7%
Nie, Nie, Nie, Nie, Tak	6	2.0%
Nie, Nie, Tak, Nie, Nie	8	2.7%
Nie, Nie, Tak, Nie, Tak	36	12.2%
Nie, Nie, Tak, Tak, Nie	2	0.7%
Nie, Nie, Tak, Tak, Tak	2	0.7%
Nie, Tak, Nie, Nie, Tak	8	2.7%
Nie, Tak, Nie, Tak, Nie	3	1.0%
Nie, Tak, Tak, Nie, Nie	43	14.6%
Nie, Tak, Tak, Nie, Tak (correct)	150	51.0%
Nie, Tak, Tak, Tak, Nie	3	1.0%
Nie, Tak, Tak, Tak, Tak	9	3.1%
Tak, Nie, Nie, Nie, Tak	2	0.7%
Tak, Nie, Nie, Tak, Nie	1	0.3%
Tak, Nie, Nie, Tak, Tak	2	0.7%
Tak, Nie, Tak, Nie, Nie	2	0.7%
Tak, Nie, Tak, Nie, Tak	5	1.7%
Tak, Nie, Tak, Tak, Nie	1	0.3%
Tak, Nie, Tak, Tak, Tak	1	0.3%
Tak, Tak, Nie, Tak, Nie	1	0.3%
Tak, Tak, Tak, Nie, Nie	1	0.3%
Tak, Tak, Tak, Tak, Tak	6	2.0%
Overall	294	100.0%

10.2.14.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Nie, Nie, Nie, Nie, Nie	1	1	0	2	2.4%	1.2%	0.0%	1.4%
Nie, Nie, Nie, Nie, Tak	1	2	3	6	2.4%	2.4%	15.0%	4.2%
Nie, Nie, Tak, Nie, Nie	2	6	0	8	4.9%	7.2%	0.0%	5.6%
Nie, Nie, Tak, Nie, Tak	6	24	6	36	14.6%	28.9%	30.0%	25.0%
Nie, Nie, Tak, Tak, Nie	2	0	0	2	4.9%	0.0%	0.0%	1.4%
Nie, Tak, Nie, Nie, Tak	3	5	0	8	7.3%	6.0%	0.0%	5.6%
Nie, Tak, Nie, Tak, Nie	1	1	1	3	2.4%	1.2%	5.0%	2.1%
Nie, Tak, Tak, Nie, Nie	8	27	8	43	19.5%	32.5%	40.0%	29.9%
Nie, Tak, Tak, Tak, Tak	3	5	1	9	7.3%	6.0%	5.0%	6.2%
Tak, Nie, Nie, Tak, Nie	1	0	0	1	2.4%	0.0%	0.0%	0.7%
Tak, Nie, Nie, Tak, Tak	2	0	0	2	4.9%	0.0%	0.0%	1.4%

Answer	Frequency				Percent			
	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Tak, Nie, Tak, Nie, Nie	1	1	0	2	2.4%	1.2%	0.0%	1.4%
Tak, Nie, Tak, Nie, Tak	3	2	0	5	7.3%	2.4%	0.0%	3.5%
Tak, Nie, Tak, Tak, Nie	1	0	0	1	2.4%	0.0%	0.0%	0.7%
Tak, Nie, Tak, Tak, Tak	1	0	0	1	2.4%	0.0%	0.0%	0.7%
Tak, Tak, Nie, Tak, Nie	1	0	0	1	2.4%	0.0%	0.0%	0.7%
Tak, Tak, Tak, Nie, Nie	1	0	0	1	2.4%	0.0%	0.0%	0.7%
Tak, Tak, Tak, Tak, Tak	3	2	1	6	7.3%	2.4%	5.0%	4.2%
Overall	41	83	20	144	100.0%	100.0%	100.0%	100.0%
Nie, Nie, Tak, Tak, Tak	0	2	0	2	0.0%	2.4%	0.0%	1.4%
Nie, Tak, Tak, Tak, Nie	0	3	0	3	0.0%	3.6%	0.0%	2.1%
Tak, Nie, Nie, Nie, Tak	0	2	0	2	0.0%	2.4%	0.0%	1.4%

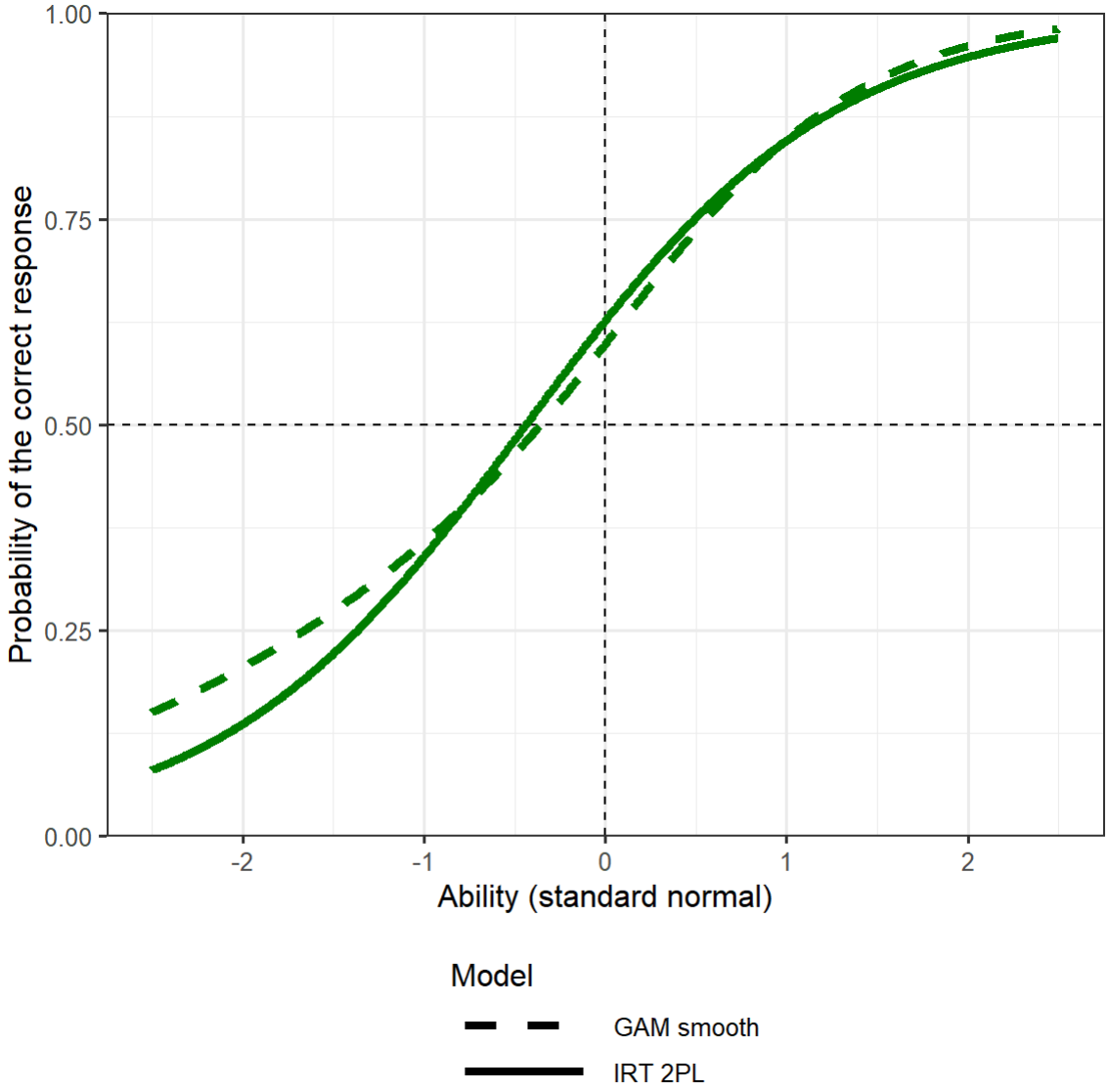
10.2.14.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	6.7	20.5	35.9	53.5	220.7	3.6	0.7
Middle quartiles	8.0	23.9	32.8	45.4	113.6	3.5	0.5
Upper quartile	10.2	20.5	28.1	37.7	97.8	3.4	0.5
Overall	6.7	22.0	31.5	46.5	220.7	3.5	0.5

10.2.15 S051110 (LLM)

10.2.15.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	61.22%	294	TIMSS	MC	0.43	0.47	1.18	−0.44	0.000

10.2.15.2 Gender DIF:

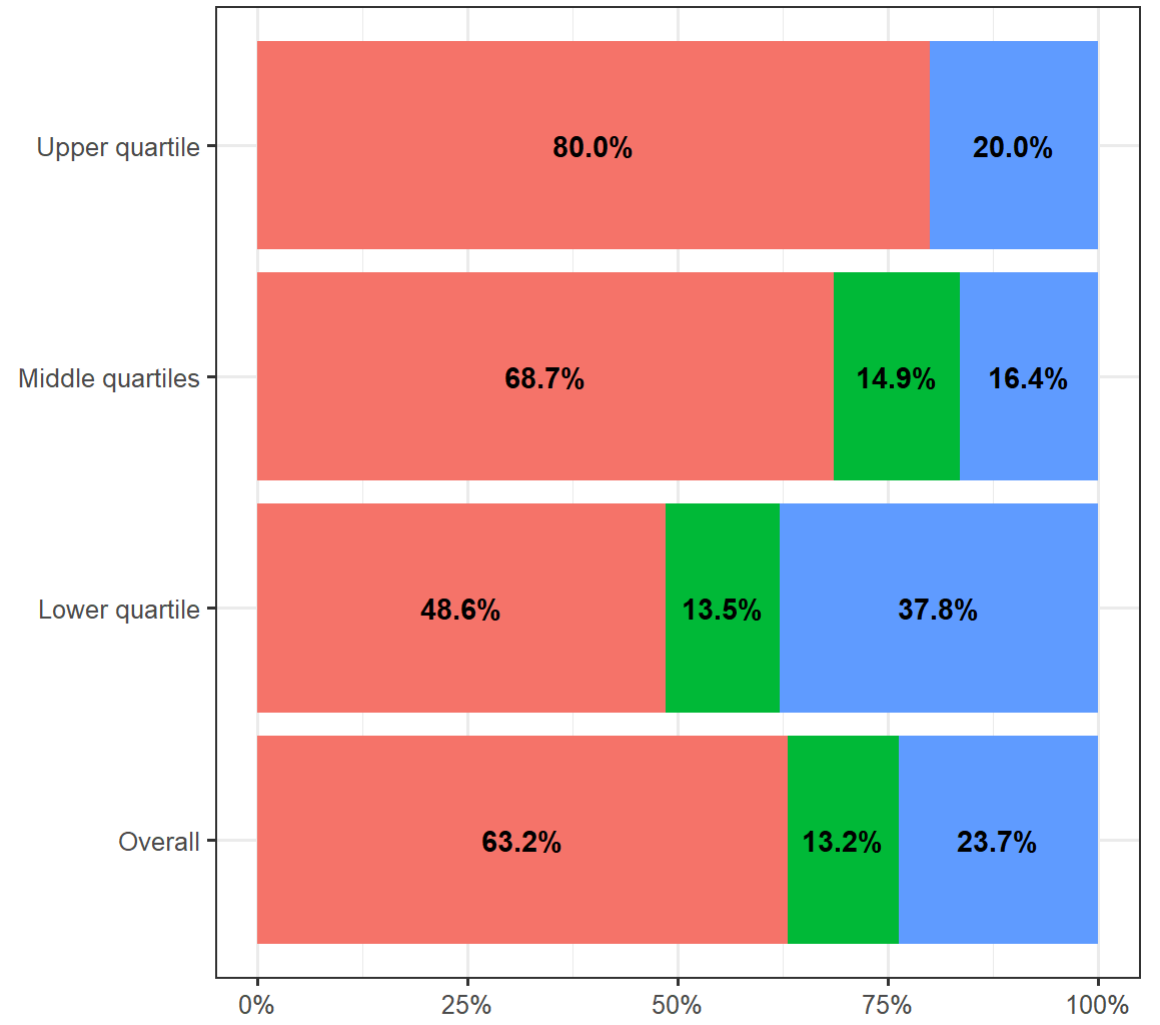
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.31	0.191	1.000	0.010	N/A

10.2.15.3 Frequency of answers:

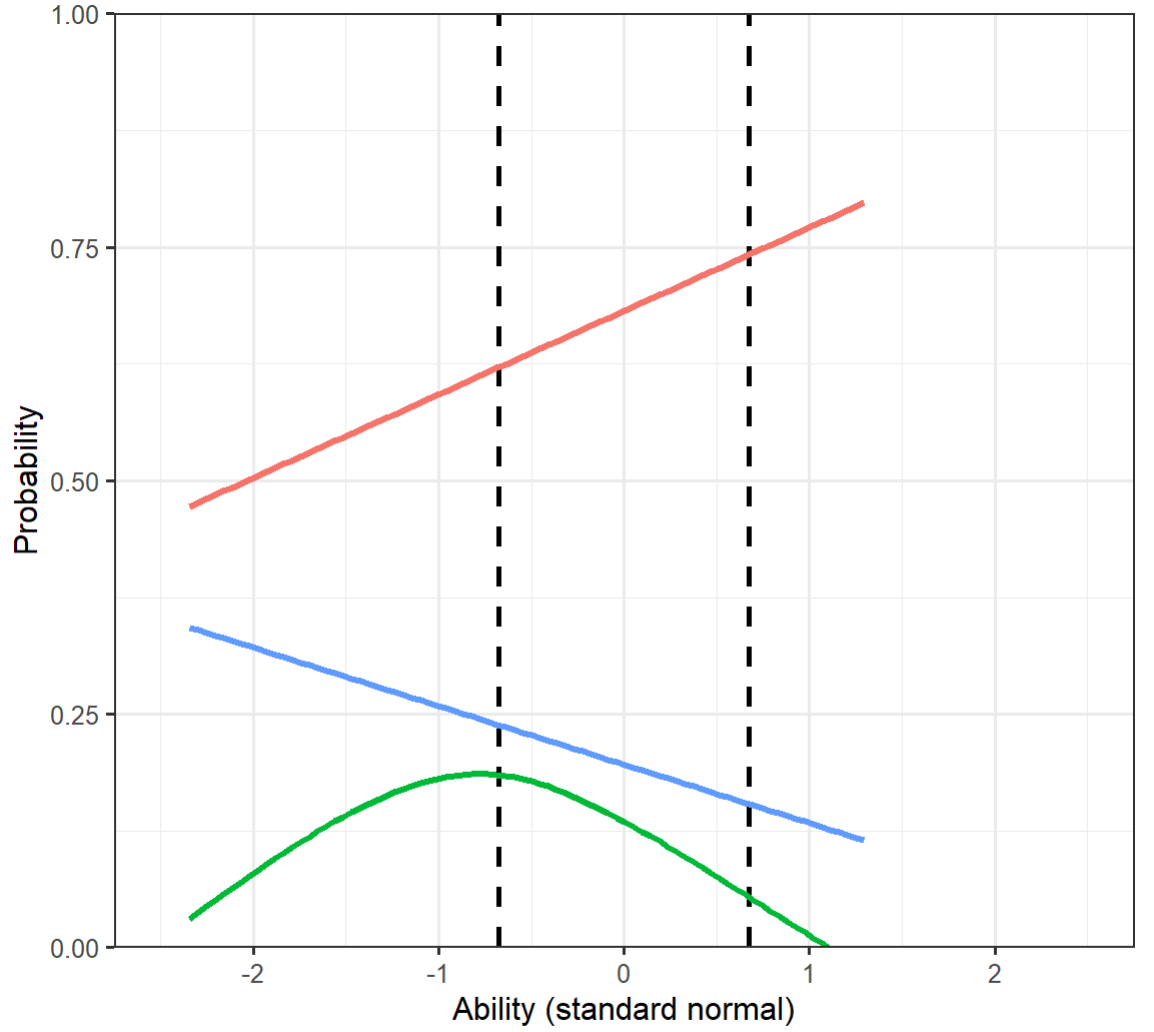
Answer	Frequency	Percent
Skorupiaki były stare, gdy umierały.	72	24.5%
Skorupiaki zjadały inne skorupiaki.	15	5.1%
Ta okolica była kiedyś pod wodą. (correct)	180	61.2%
W tej okolicy żyły różne gatunki roślin.	27	9.2%
Overall	294	100.0%

10.2.15.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Skorupiaki były stare, gdy umierały.	18	46	8	72	48.6%	68.7%	80.0%	63.2%
Skorupiaki zjadały inne skorupiaki.	5	10	0	15	13.5%	14.9%	0.0%	13.2%
W tej okolicy żyły różne gatunki roślin.	14	11	2	27	37.8%	16.4%	20.0%	23.7%
Overall	37	67	10	114	100.0%	100.0%	100.0%	100.0%

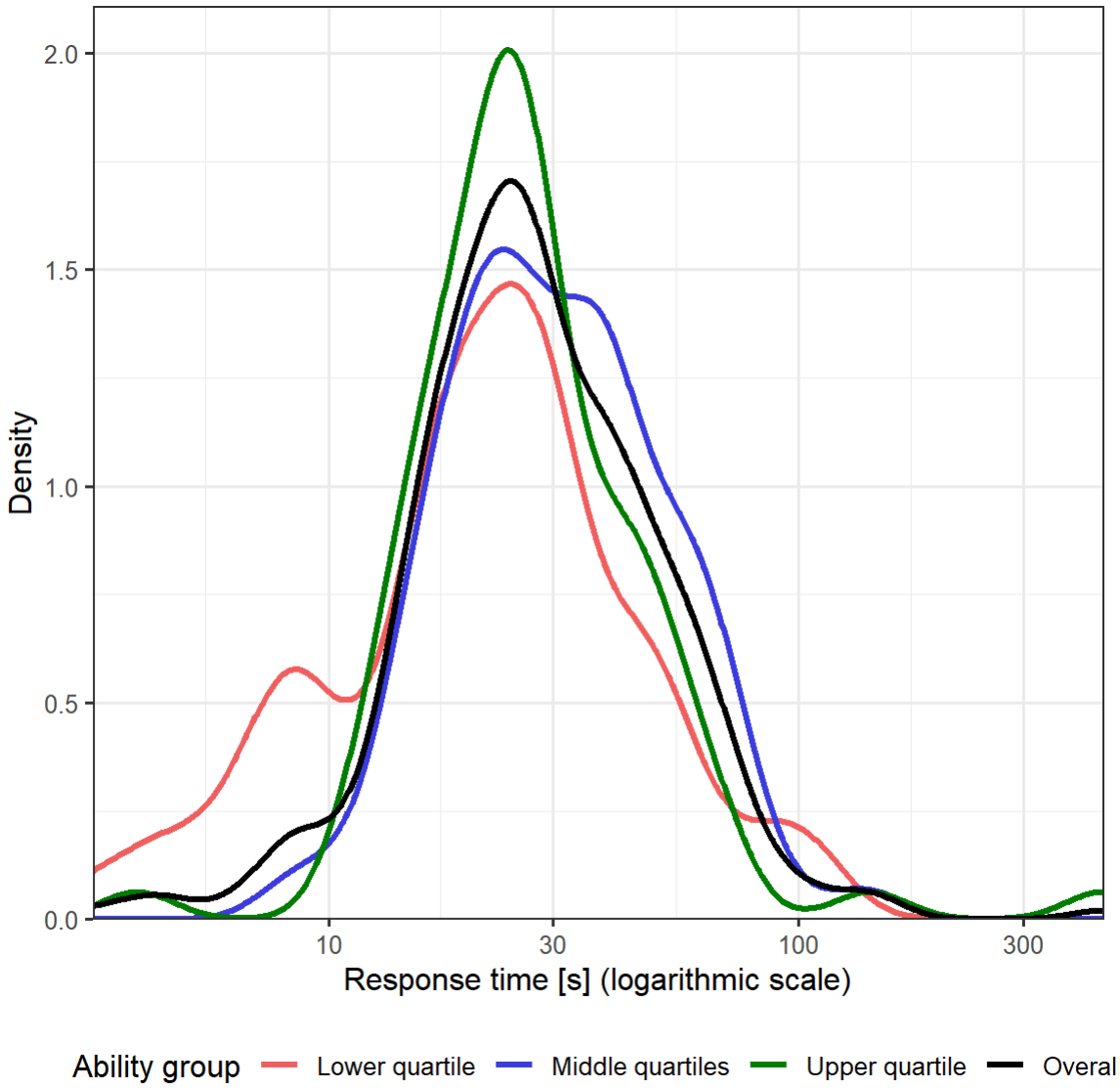



```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



korupiaki były stare, gdy umierały. Skorupiaki zjadały inne skorupiaki. W tej okolicy żyły ró

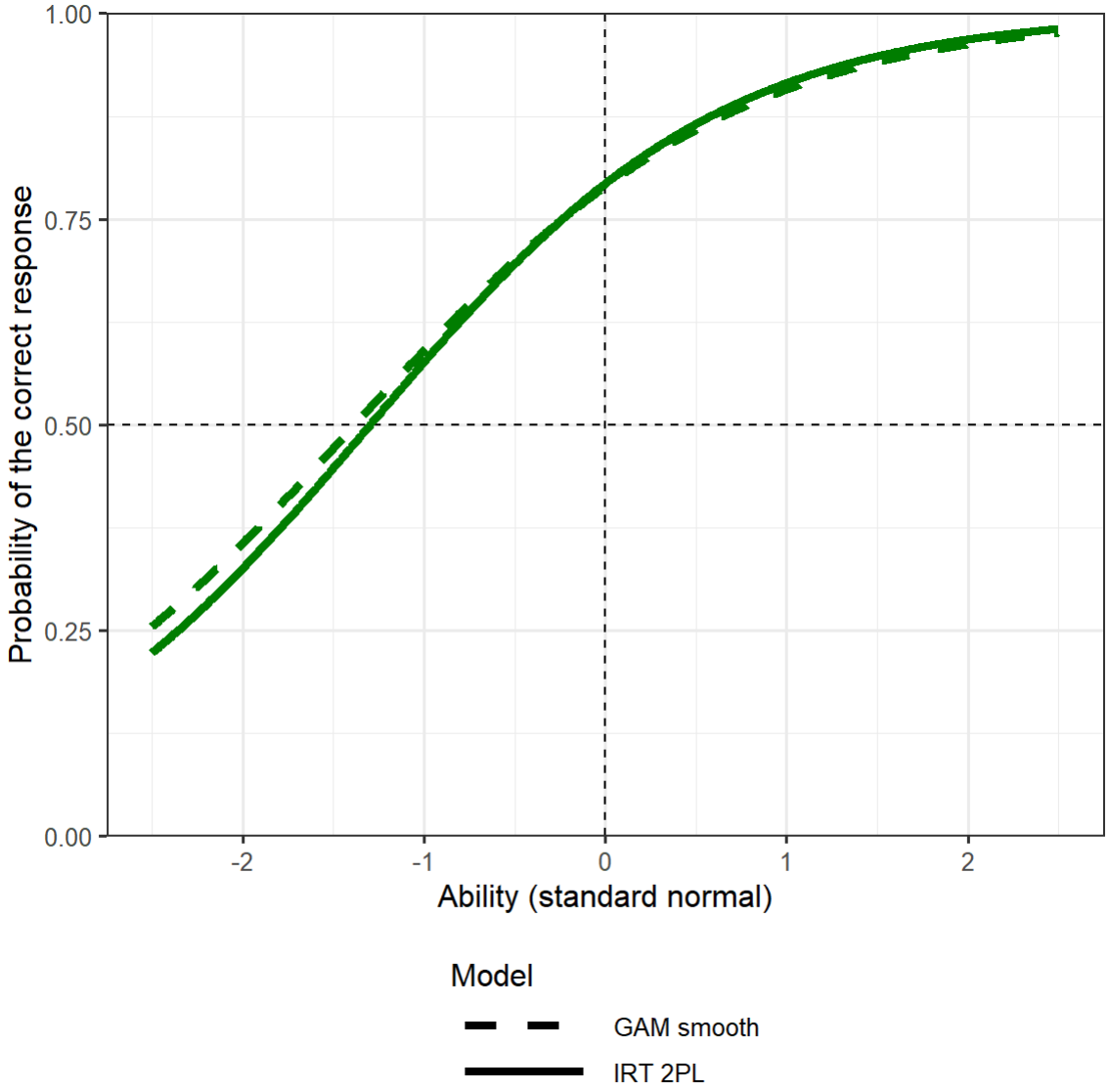
10.2.15.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	15.1	22.9	30.8	105.6	3.1	0.8
Middle quartiles	8.2	20.5	28.7	42.3	146.6	3.4	0.5
Upper quartile	3.9	19.1	24.9	36.8	444.4	3.3	0.6
Overall	3.1	19.1	26.4	40.0	444.4	3.3	0.6

10.2.16 S090080 (LLM)

10.2.16.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	75.00%	281	LLM teachers' selection	MC	0.37	0.23	1.04	−1.30	0.0

10.2.16.2 Gender DIF:

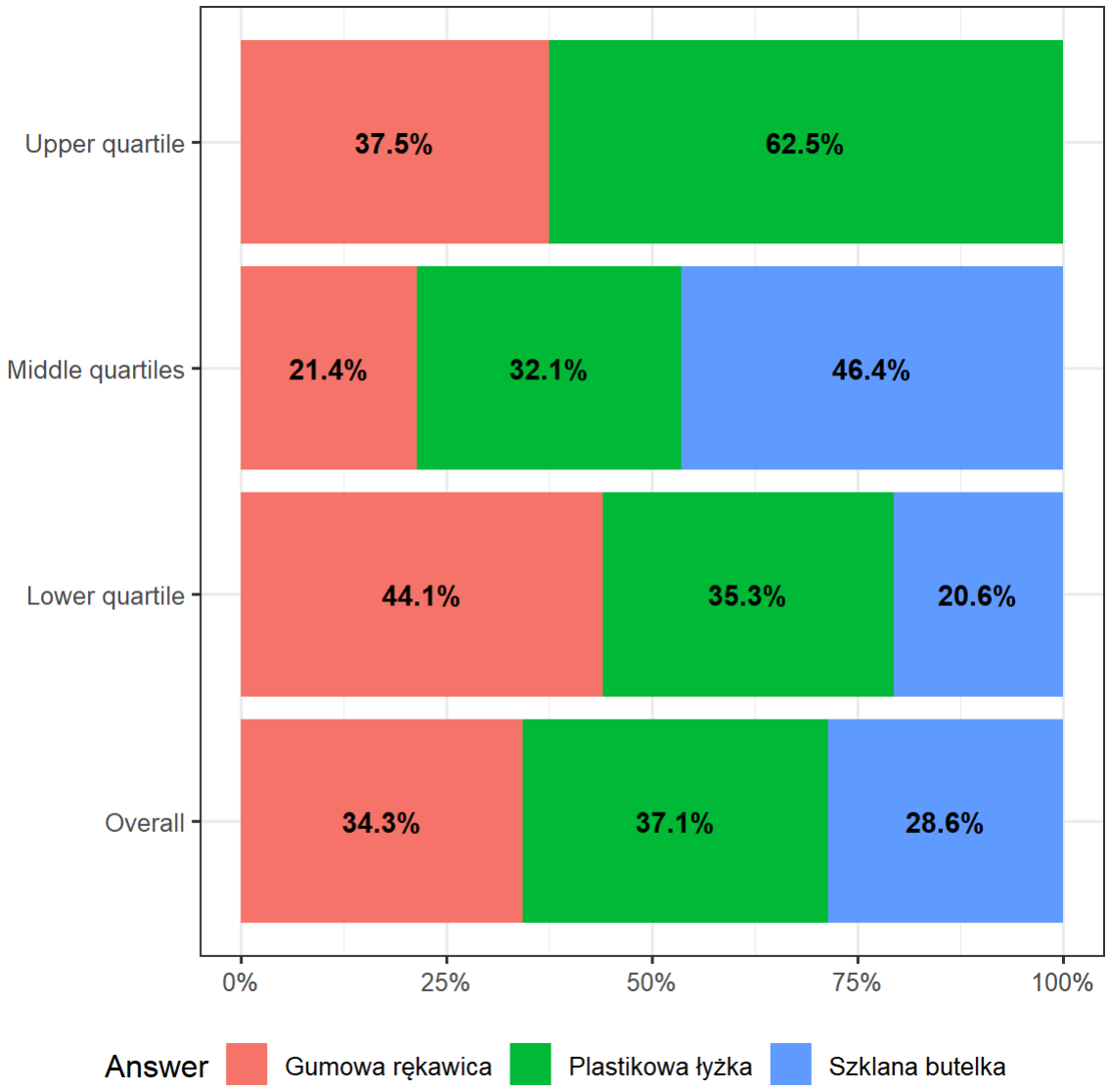
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
7.15	0.028	1.000	0.026	N/A

10.2.16.3 Frequency of answers:

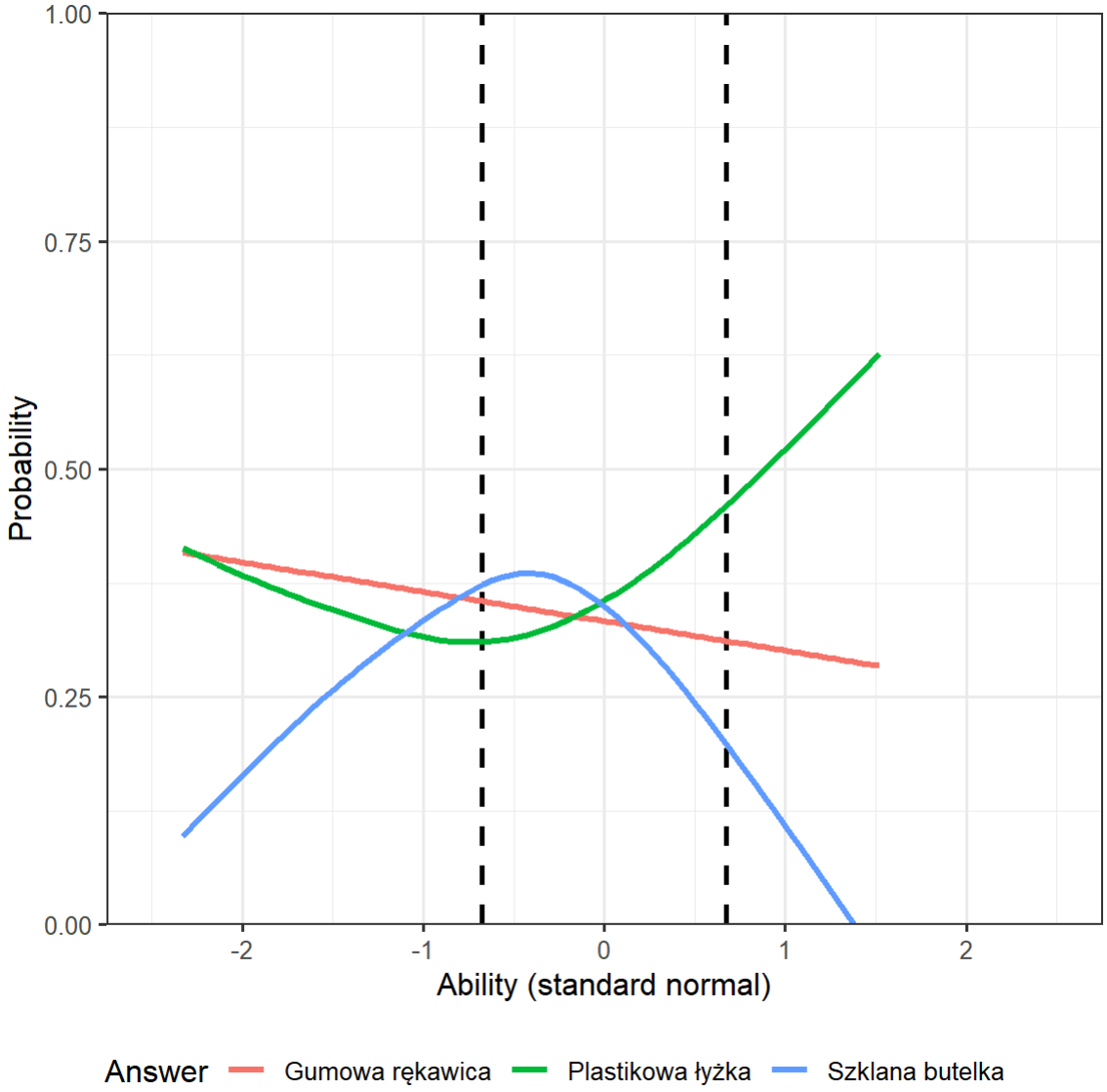
Answer	Frequency	Percent
Gumowa rękawica	24	8.6%
Plastikowa łyżka	26	9.3%
Srebrny widelec (correct)	210	75.0%
Szklana butelka	20	7.1%
Overall	280	100.0%

10.2.16.4 Distractors functioning:

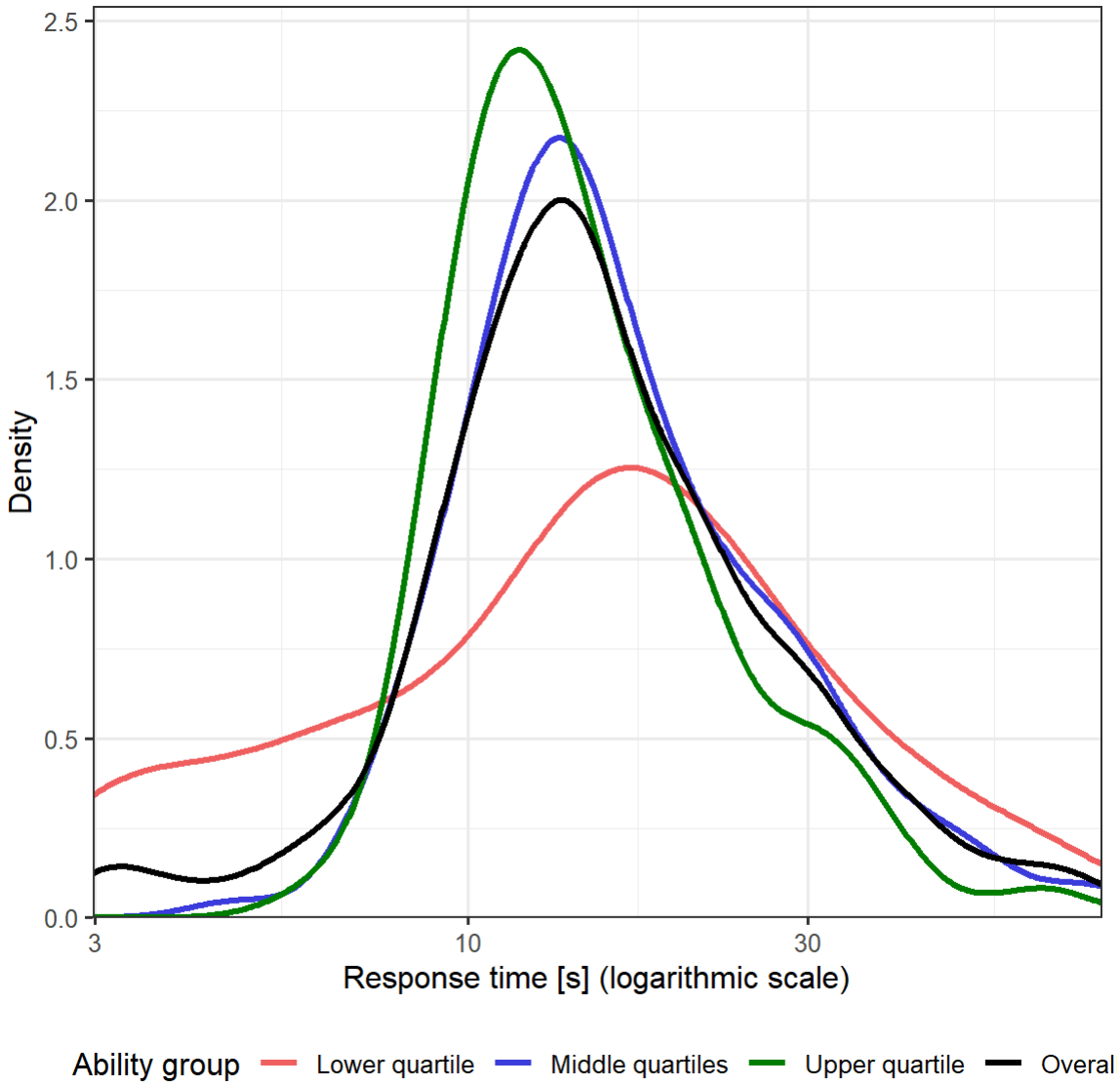
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Gumowa rękawica	15	6	3	24	44.1%	21.4%	37.5%	34.3%
Plastikowa łyżka	12	9	5	26	35.3%	32.1%	62.5%	37.1%
Szklana butelka	7	13	0	20	20.6%	46.4%	0.0%	28.6%
Overall	34	28	8	70	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



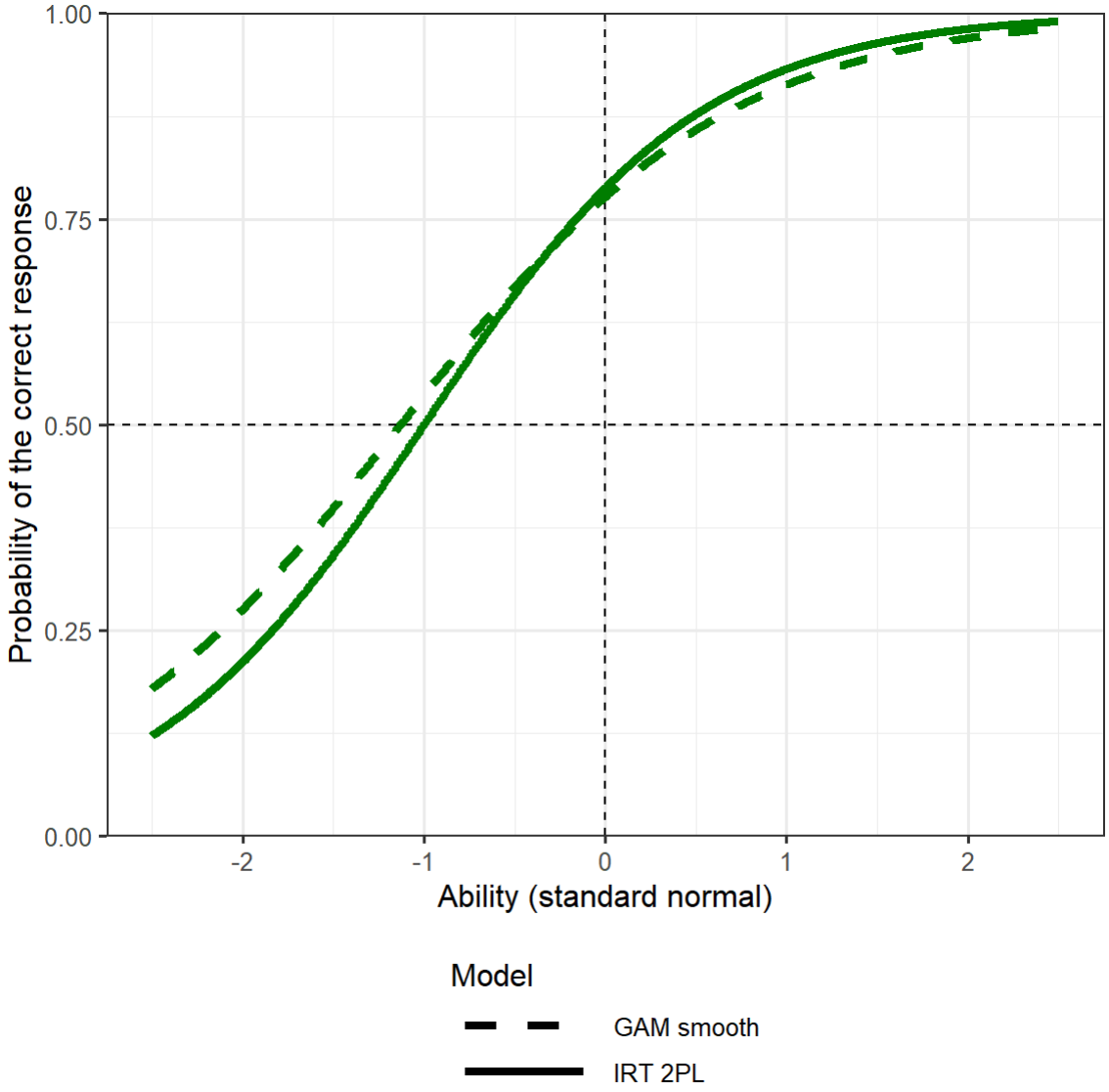
10.2.16.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	8.6	15.0	22.8	70.6	2.7	0.8
Middle quartiles	4.7	11.7	14.7	21.2	77.3	2.8	0.5
Upper quartile	6.3	10.8	13.4	19.1	64.3	2.7	0.4
Overall	3.0	11.0	14.5	21.1	77.3	2.7	0.6

10.2.17 SP71140 (LLM)

10.2.17.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	73.83%	299	TIMSS	MC	0.40	0.44	1.31	−1.00	0.000

10.2.17.2 Gender DIF:

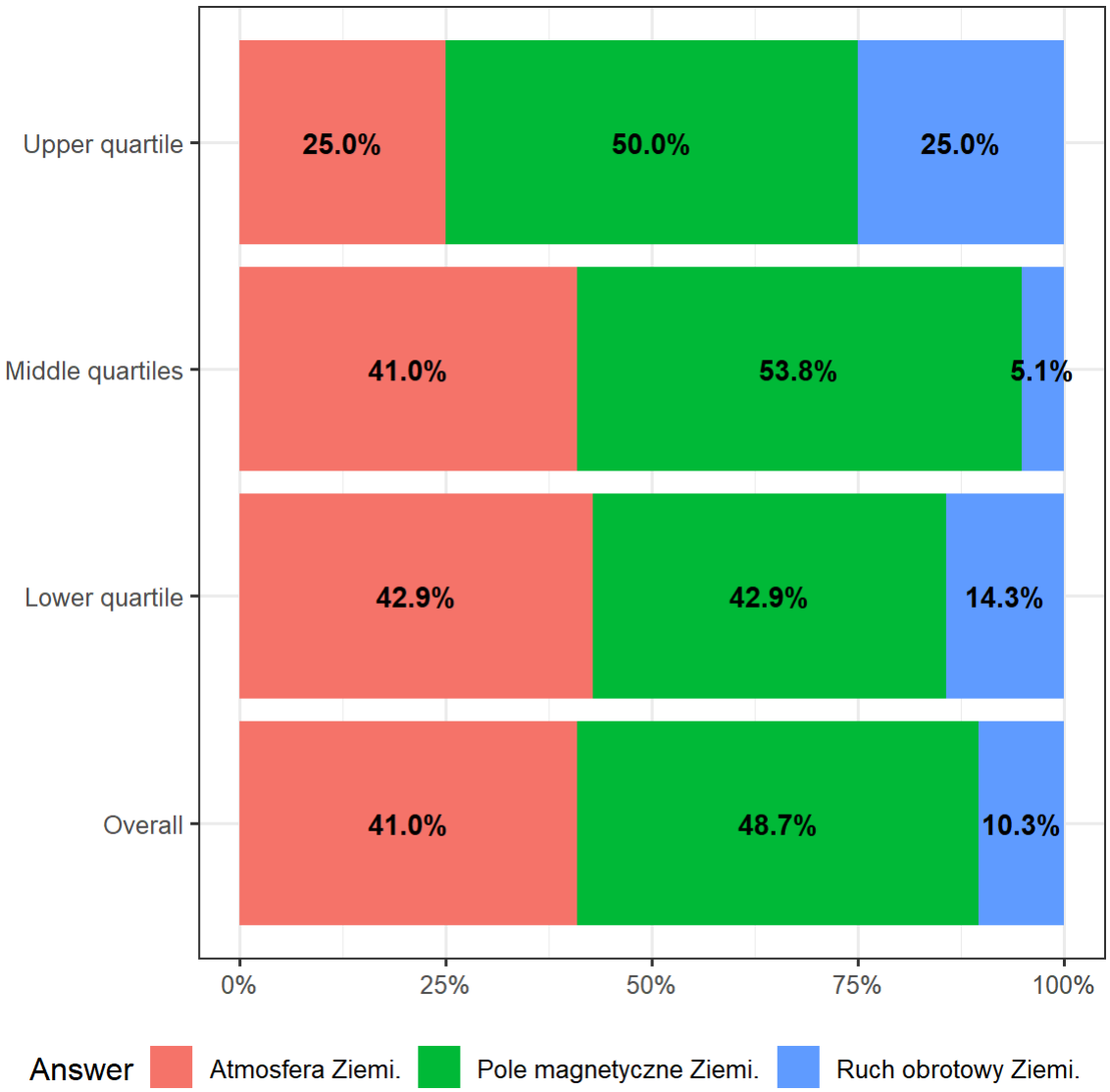
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.11	0.575	1.000	0.004	N/A

10.2.17.3 Frequency of answers:

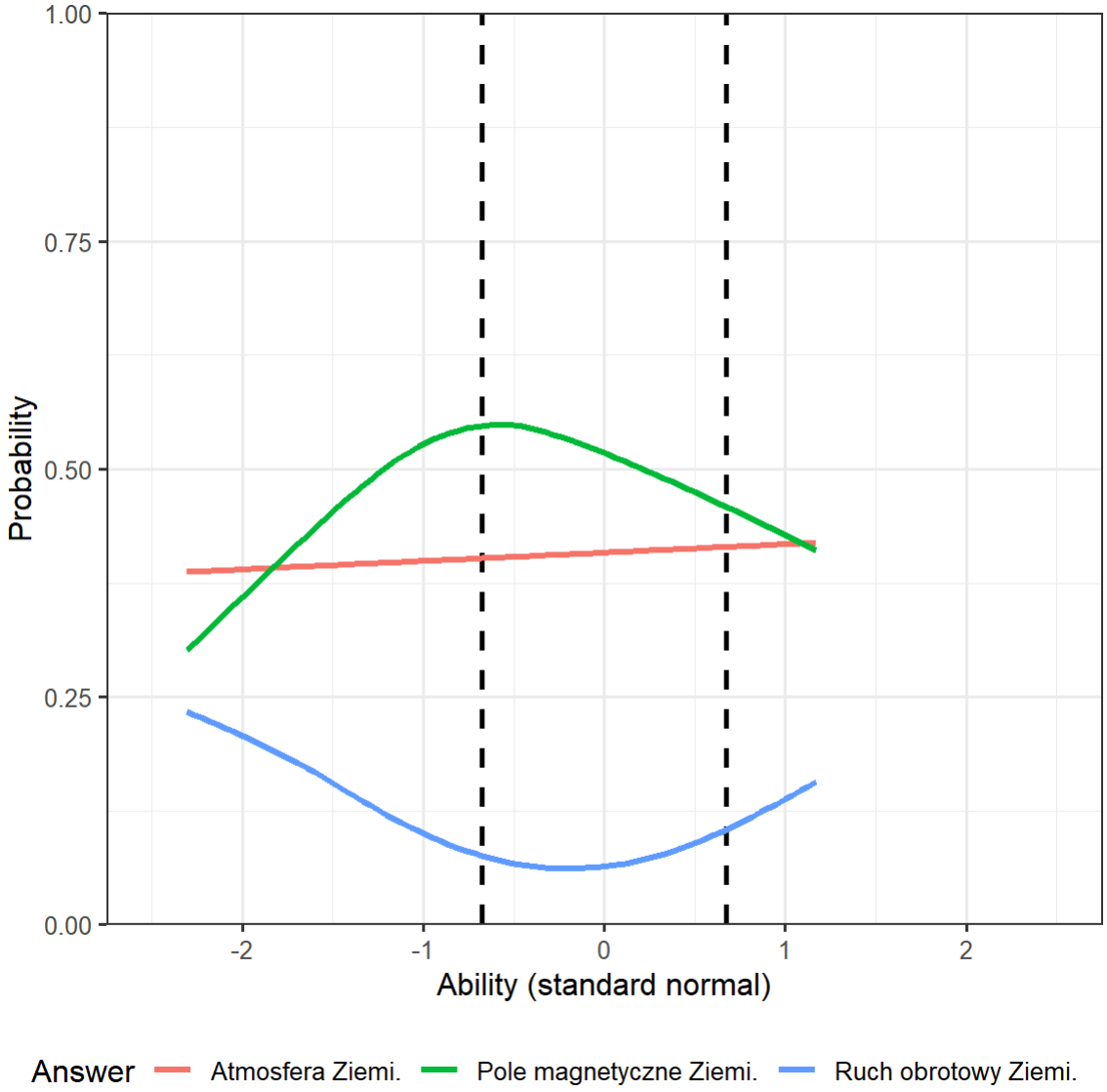
Answer	Frequency	Percent
Atmosfera Ziemi.	32	10.7%
Grawitacja Ziemi. (correct)	220	73.8%
Pole magnetyczne Ziemi.	38	12.8%
Ruch obrotowy Ziemi.	8	2.7%
Overall	298	100.0%

10.2.17.4 Distractors functioning:

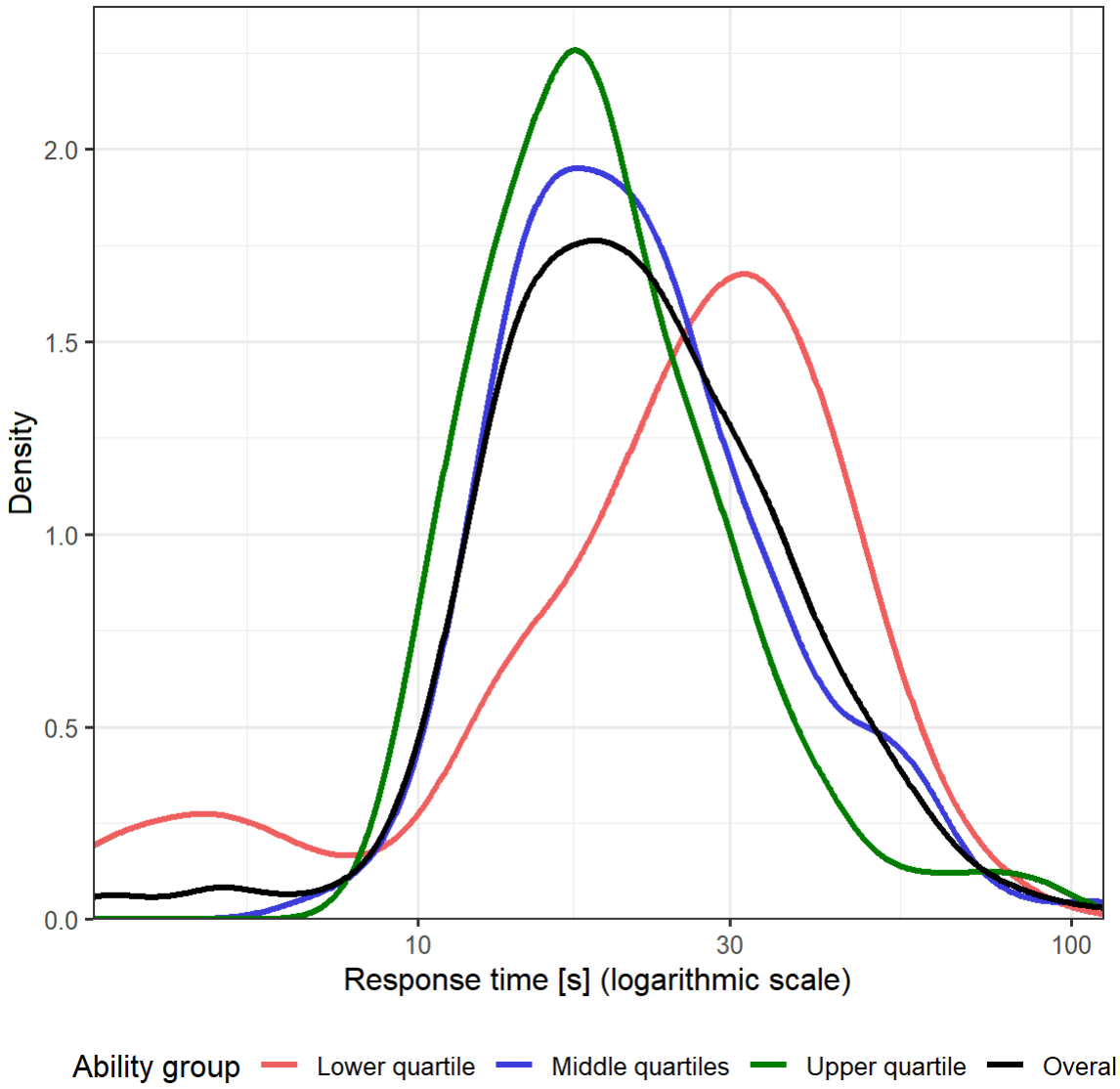
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Atmosfera Ziemi.	15	16	1	32	42.9%	41.0%	25.0%	41.0%
Pole magnetyczne Ziemi.	15	21	2	38	42.9%	53.8%	50.0%	48.7%
Ruch obrotowy Ziemi.	5	2	1	8	14.3%	5.1%	25.0%	10.3%
Overall	35	39	4	78	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



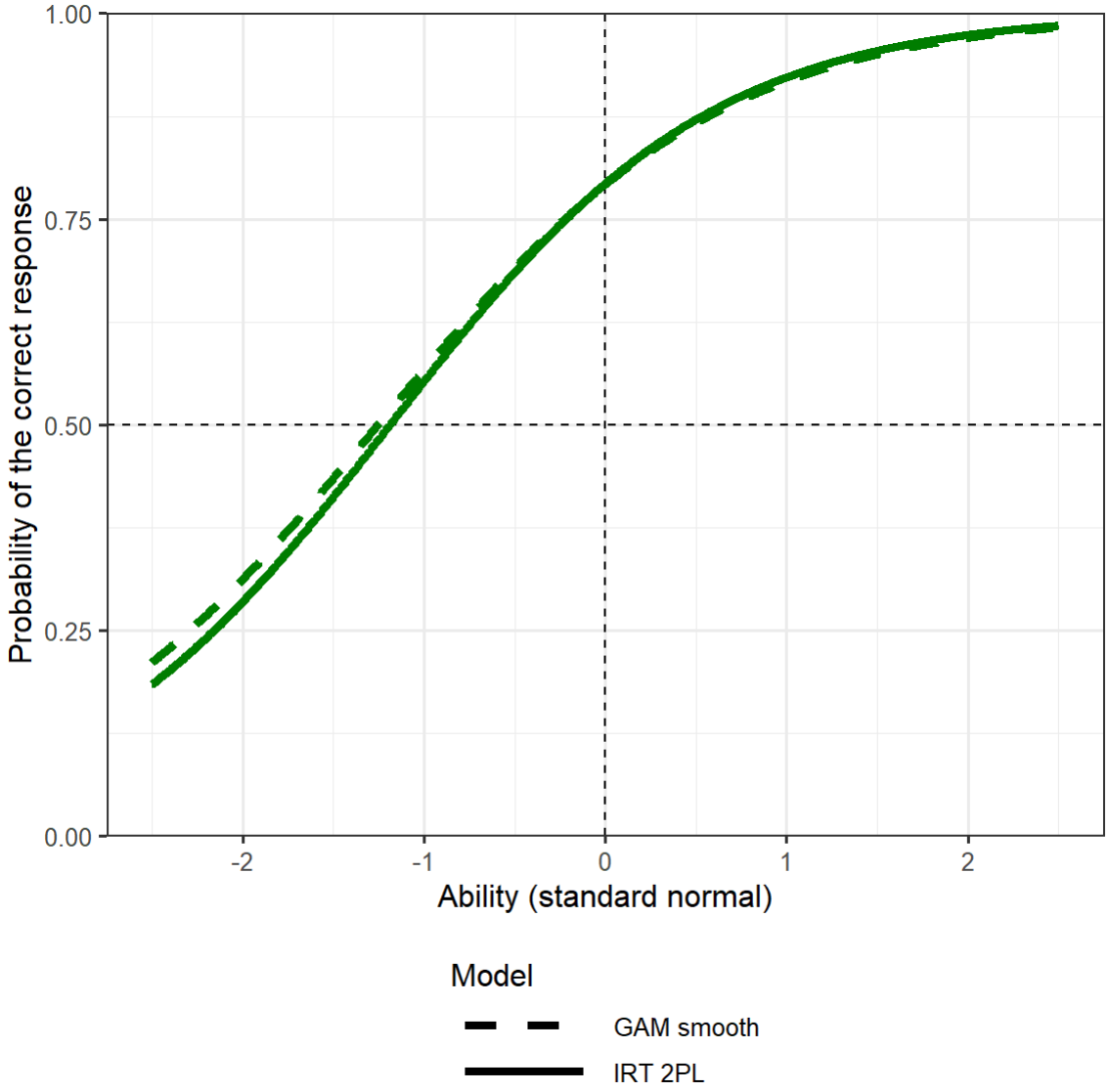
10.2.17.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.2	16.0	26.6	36.0	74.0	3.1	0.7
Middle quartiles	7.0	15.4	20.1	28.1	112.0	3.1	0.5
Upper quartile	9.6	14.0	18.4	24.6	89.7	3.0	0.4
Overall	3.2	15.1	20.9	30.1	112.0	3.1	0.5

10.2.18 S051105 (LLM)

10.2.18.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	73.77%	306	TIMSS	MC	0.40	0.44	1.13	−1.19	0.016

10.2.18.2 Gender DIF:

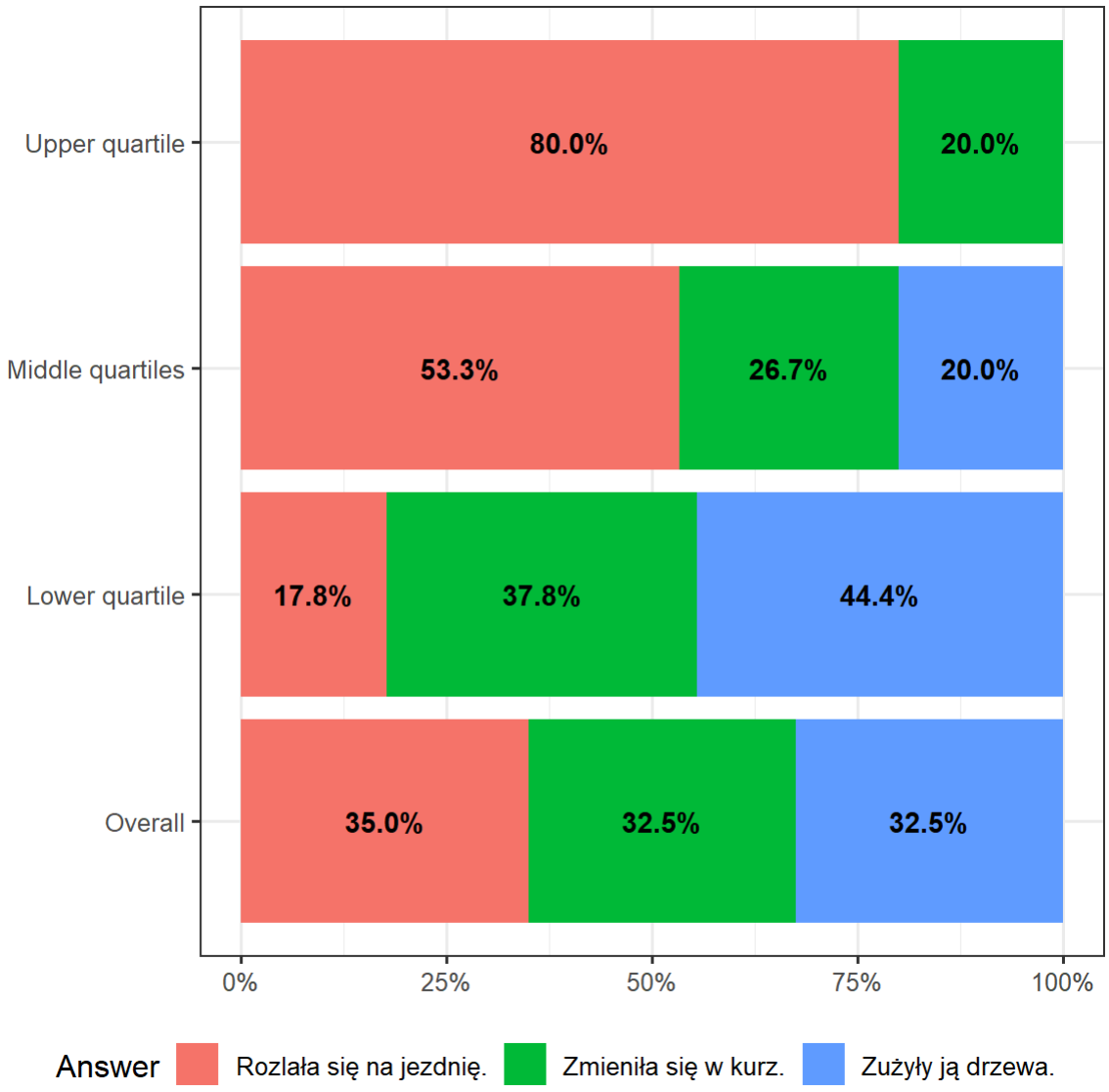
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
6.88	0.032	1.000	0.022	N/A

10.2.18.3 Frequency of answers:

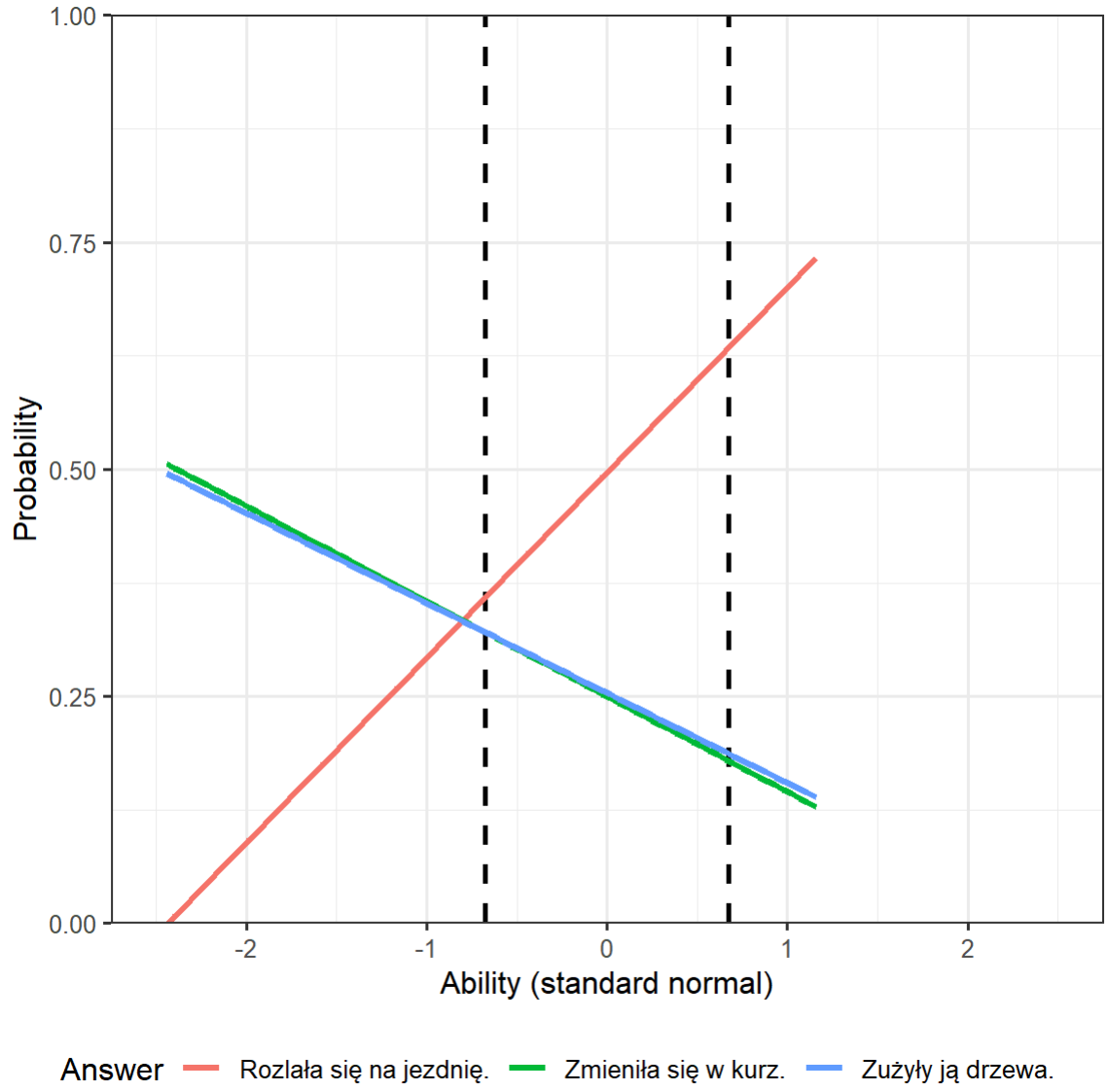
Answer	Frequency	Percent
Rozlała się na jezdnię.	28	9.2%
Uniosła się w powietrze. (correct)	225	73.8%
Zmieniła się w kurz.	26	8.5%
Zużyły ją drzewa.	26	8.5%
Overall	305	100.0%

10.2.18.4 Distractors functioning:

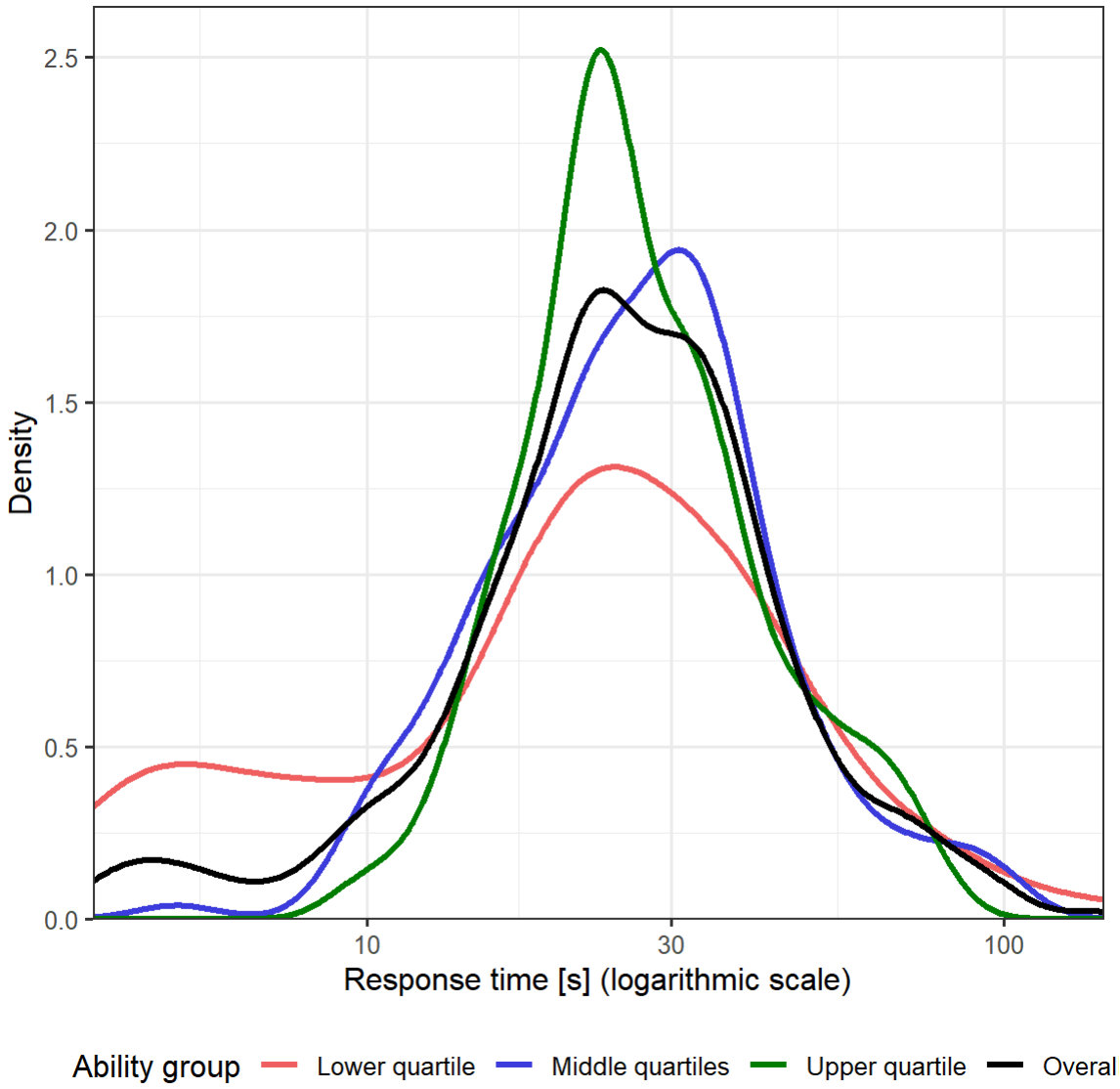
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Rozlała się na jezdnię.	8	16	4	28	17.8%	53.3%	80.0%	35.0%
Zmieniła się w kurz.	17	8	1	26	37.8%	26.7%	20.0%	32.5%
Zużyły ją drzewa.	20	6	0	26	44.4%	20.0%	0.0%	32.5%
Overall	45	30	5	80	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



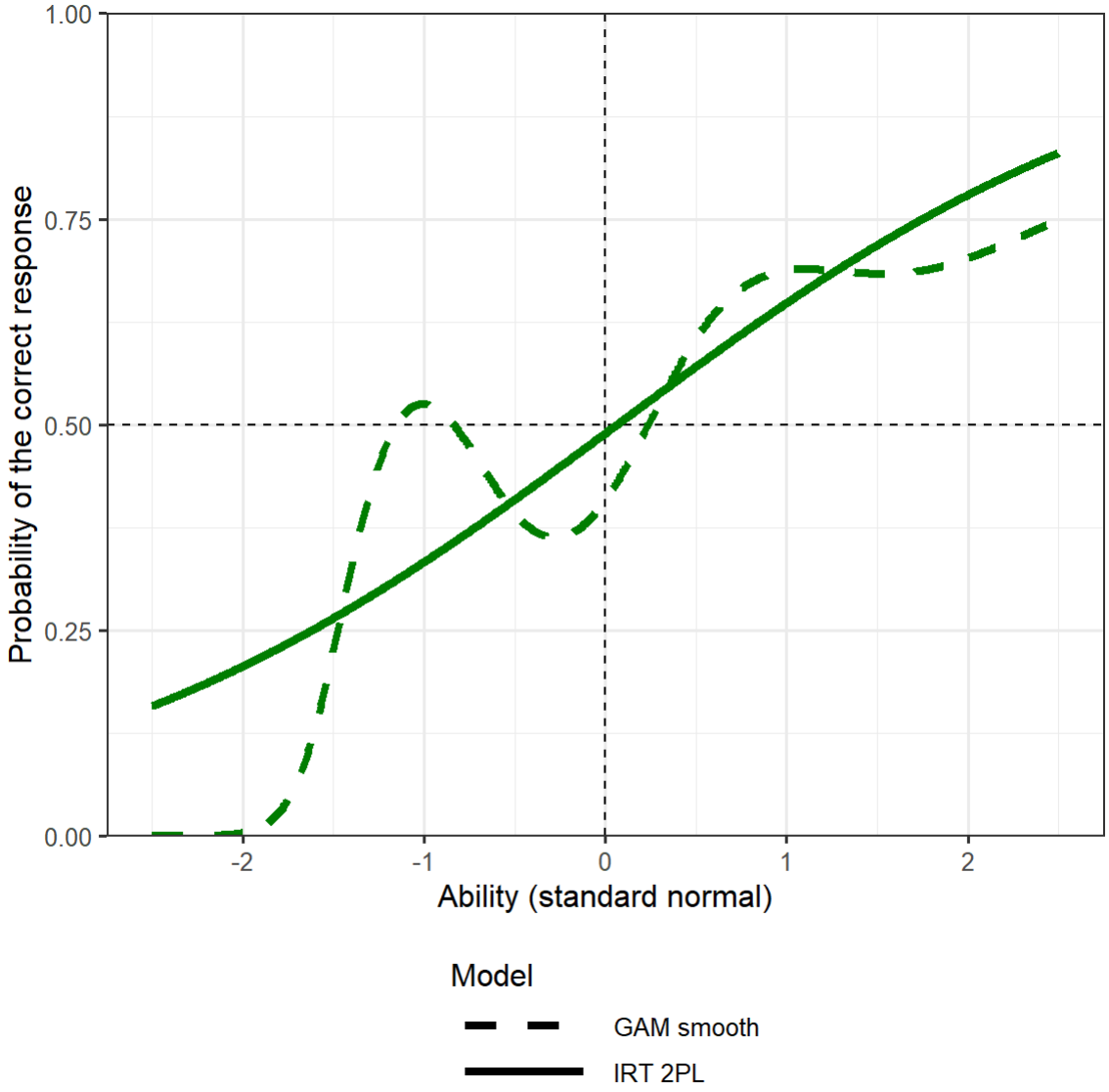
10.2.18.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.7	13.3	22.6	36.0	143.0	3.0	0.8
Middle quartiles	5.0	18.9	27.5	35.3	95.2	3.3	0.5
Upper quartile	9.9	21.2	24.6	33.6	76.5	3.3	0.4
Overall	3.7	18.2	24.8	35.5	143.0	3.2	0.6

10.2.19 S890026 (TIMSS)

10.2.19.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	48.93%	281	LLM complex setup	MC	0.30	0.12	0.65	0.07	0.02

10.2.19.2 Gender DIF:

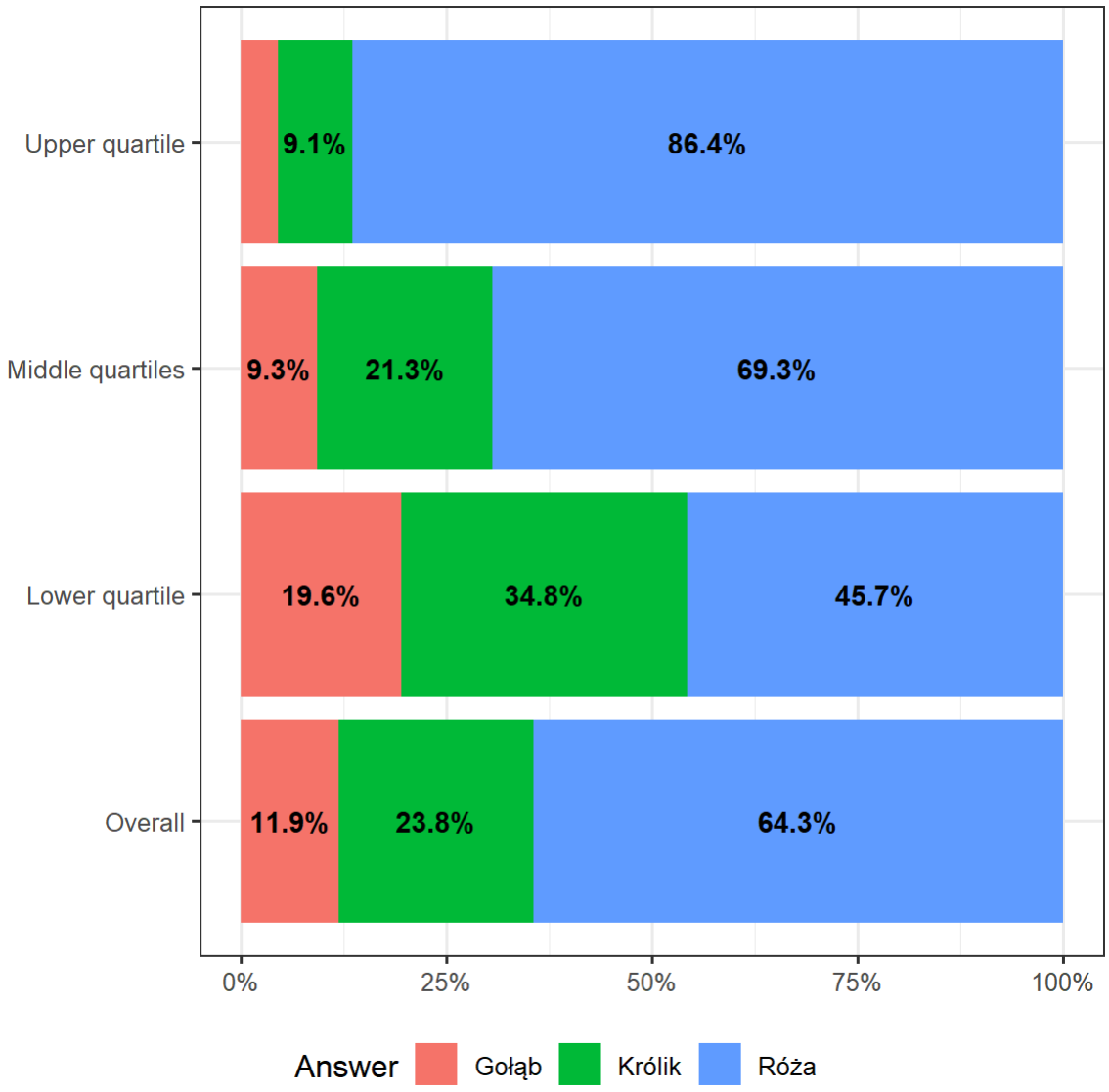
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.39	0.824	1.000	0.001	N/A

10.2.19.3 Frequency of answers:

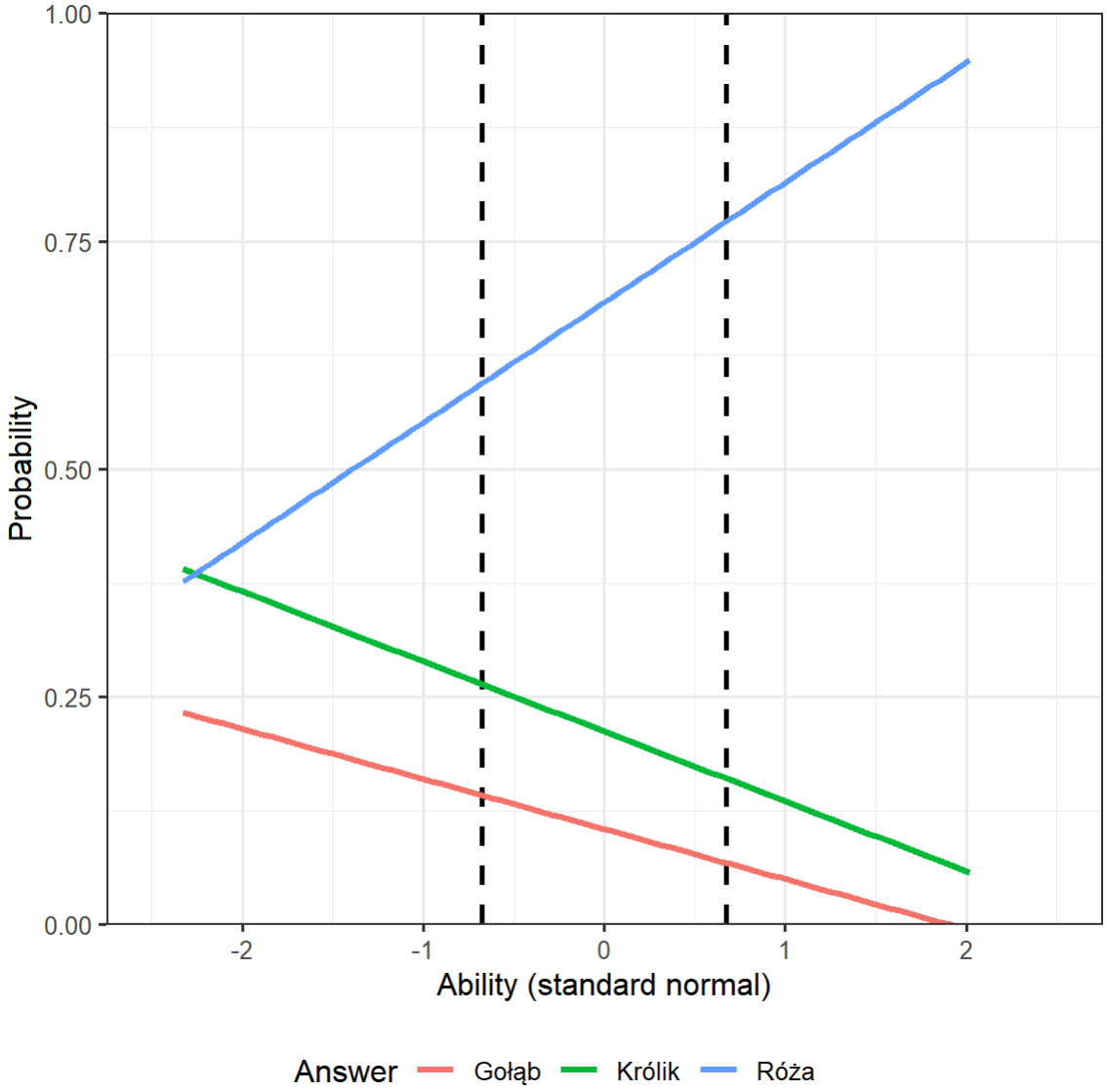
Answer	Frequency	Percent
Gołąb	17	6.1%
Królik	34	12.1%
Róża	92	32.9%
Żaba (correct)	137	48.9%
Overall	280	100.0%

10.2.19.4 Distractors functioning:

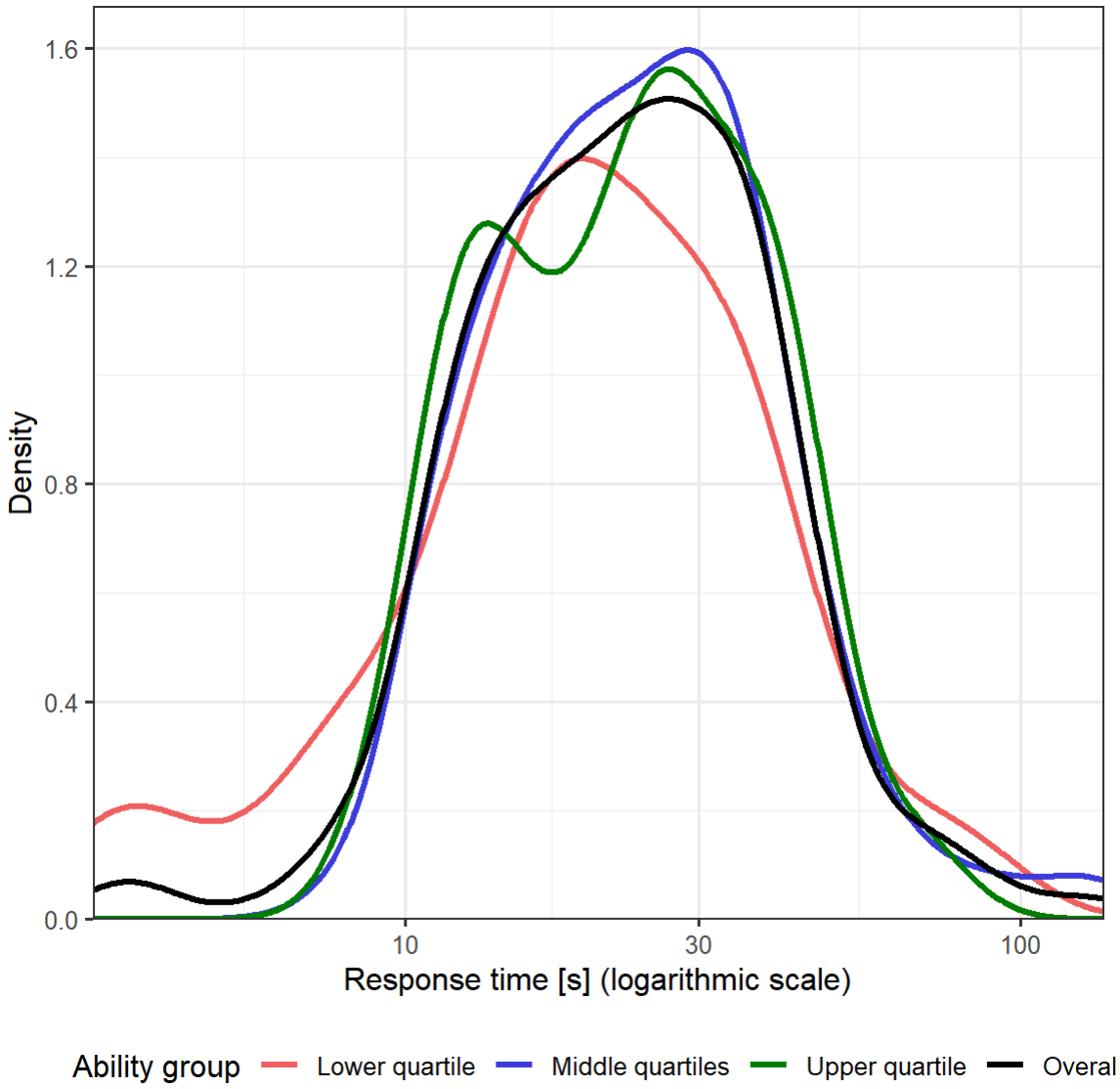
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Gołąb	9	7	1	17	19.6%	9.3%	4.5%	11.9%
Królik	16	16	2	34	34.8%	21.3%	9.1%	23.8%
Róża	21	52	19	92	45.7%	69.3%	86.4%	64.3%
Overall	46	75	22	143	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



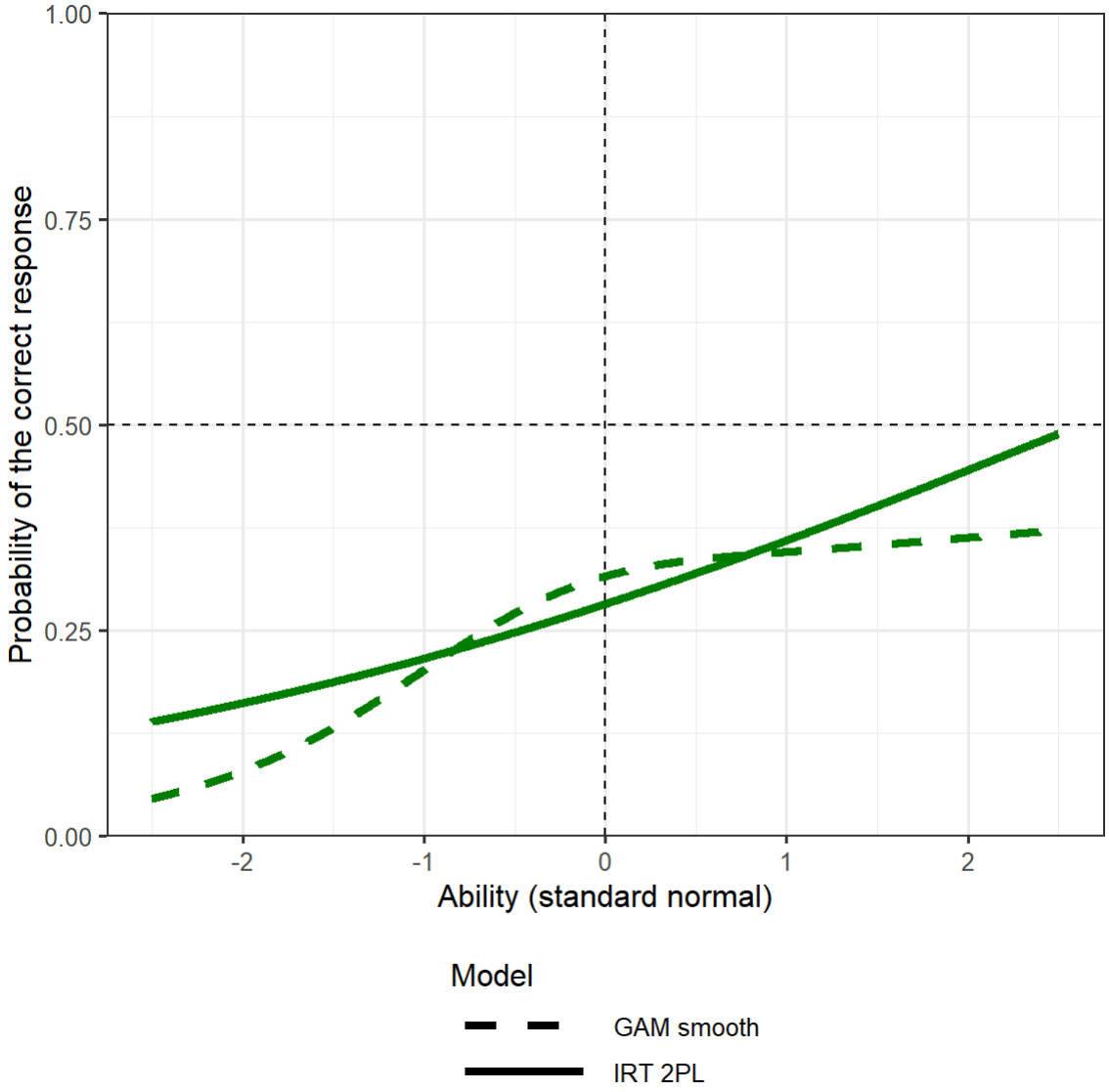
10.2.19.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	14.0	20.4	33.0	87.3	3.0	0.7
Middle quartiles	7.8	15.9	23.7	34.1	136.5	3.2	0.5
Upper quartile	9.2	14.9	23.8	34.5	68.5	3.1	0.5
Overall	3.1	15.2	23.0	33.9	136.5	3.1	0.6

10.2.20 S100015 (LLM)

10.2.20.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	28.62%	305	LLM simple setup	MC	0.15	0.16	0.36	2.62	0.000

10.2.20.2 Gender DIF:

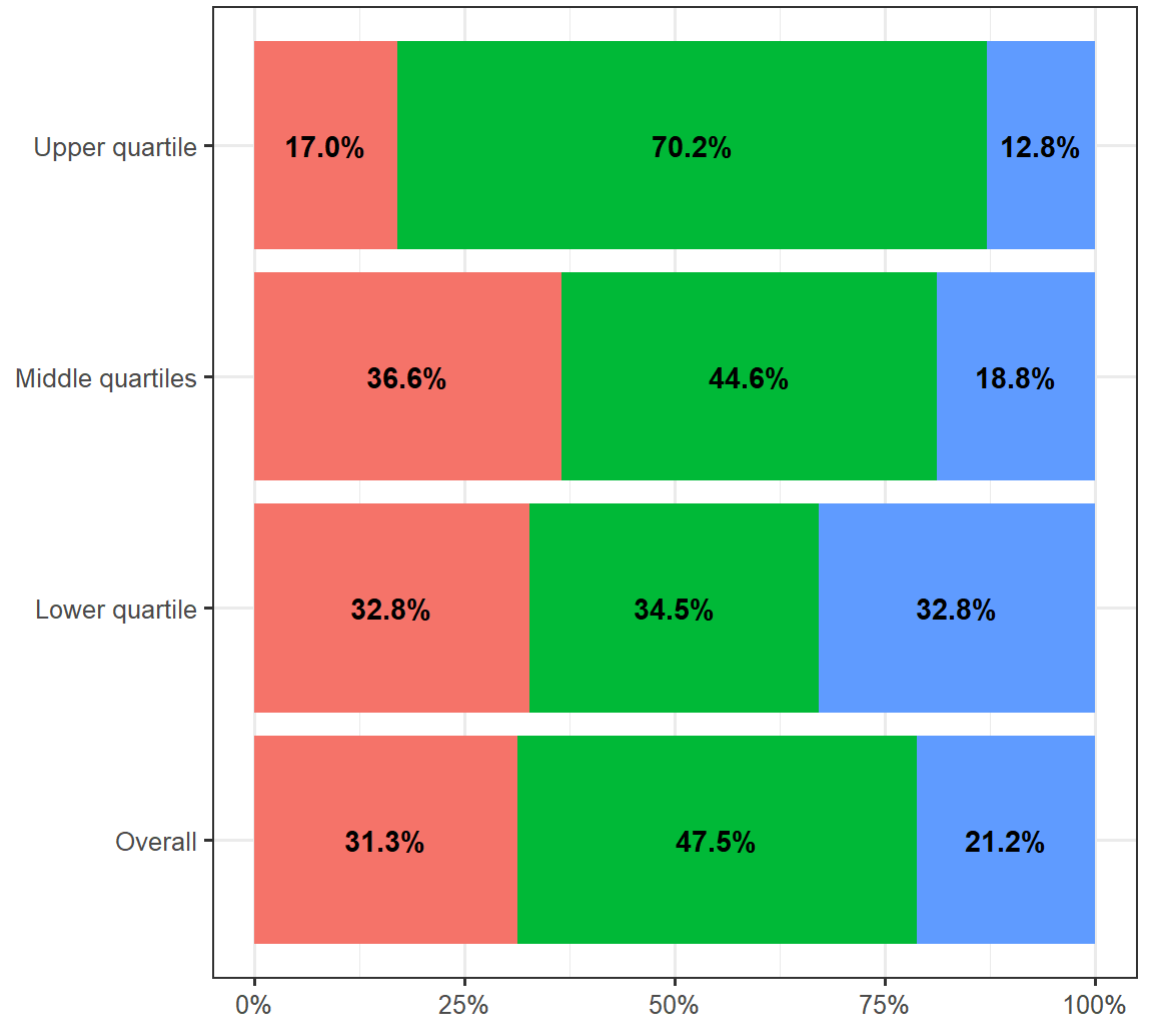
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
4.37	0.112	1.000	0.014	N/A

10.2.20.3 Frequency of answers:

Answer	Frequency	Percent
Drewno ma większą gęstość niż woda.	68	22.4%
Guma ma mniejszą gęstość niż metal, ale większą niż woda. (correct)	87	28.6%
Guma ma większą gęstość niż drewno, ale mniejszą niż metal.	103	33.9%
Metal ma mniejszą gęstość niż woda.	46	15.1%
Overall	304	100.0%

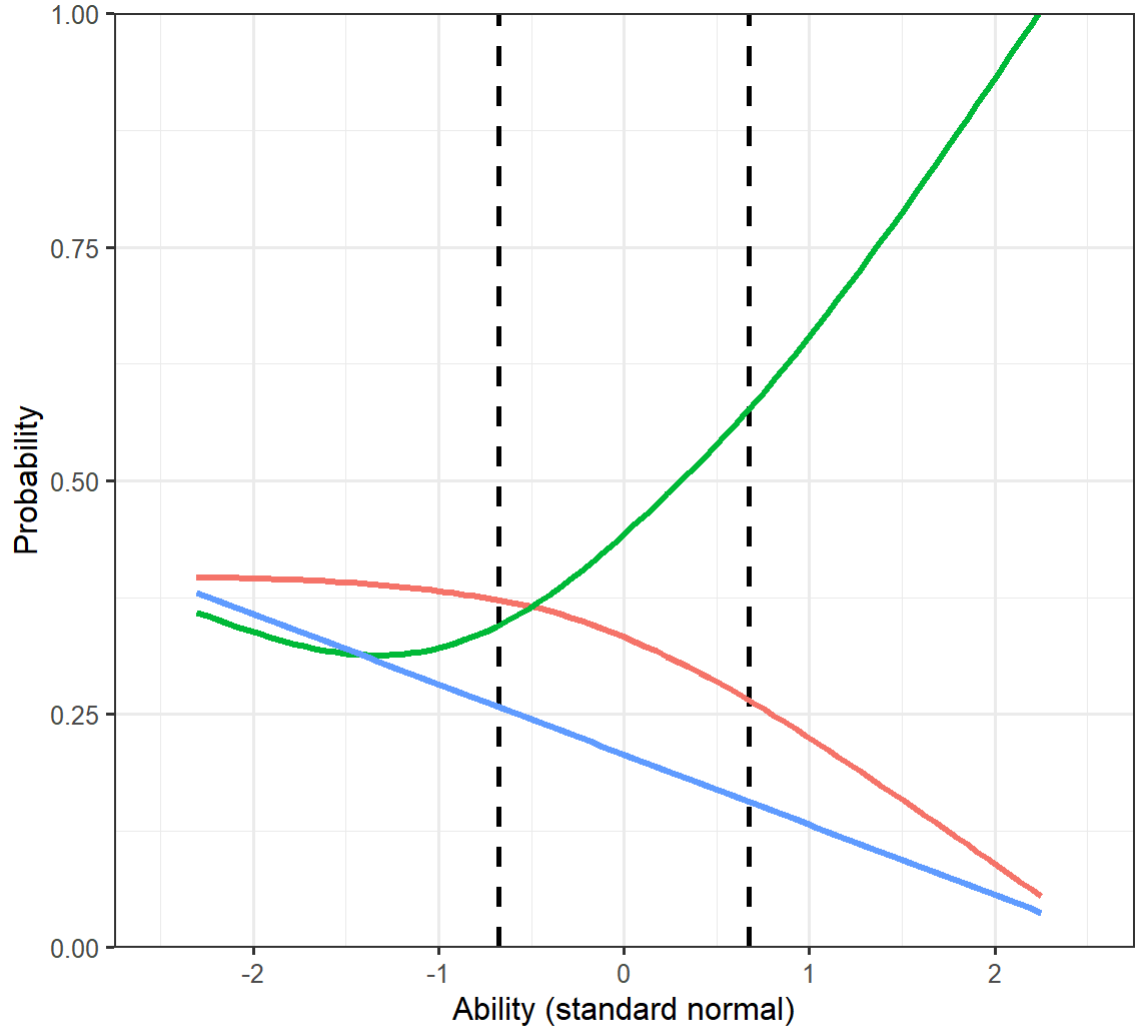
10.2.20.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Drewno ma większą gęstość niż woda.	19	41	8	68	32.8%	36.6%	17.0%	31.3%
Guma ma większą gęstość niż drewno, ale mniejszą niż metal.	20	50	33	103	34.5%	44.6%	70.2%	47.5%
Metal ma mniejszą gęstość niż woda.	19	21	6	46	32.8%	18.8%	12.8%	21.2%
Overall	58	112	47	217	100.0%	100.0%	100.0%	100.0%



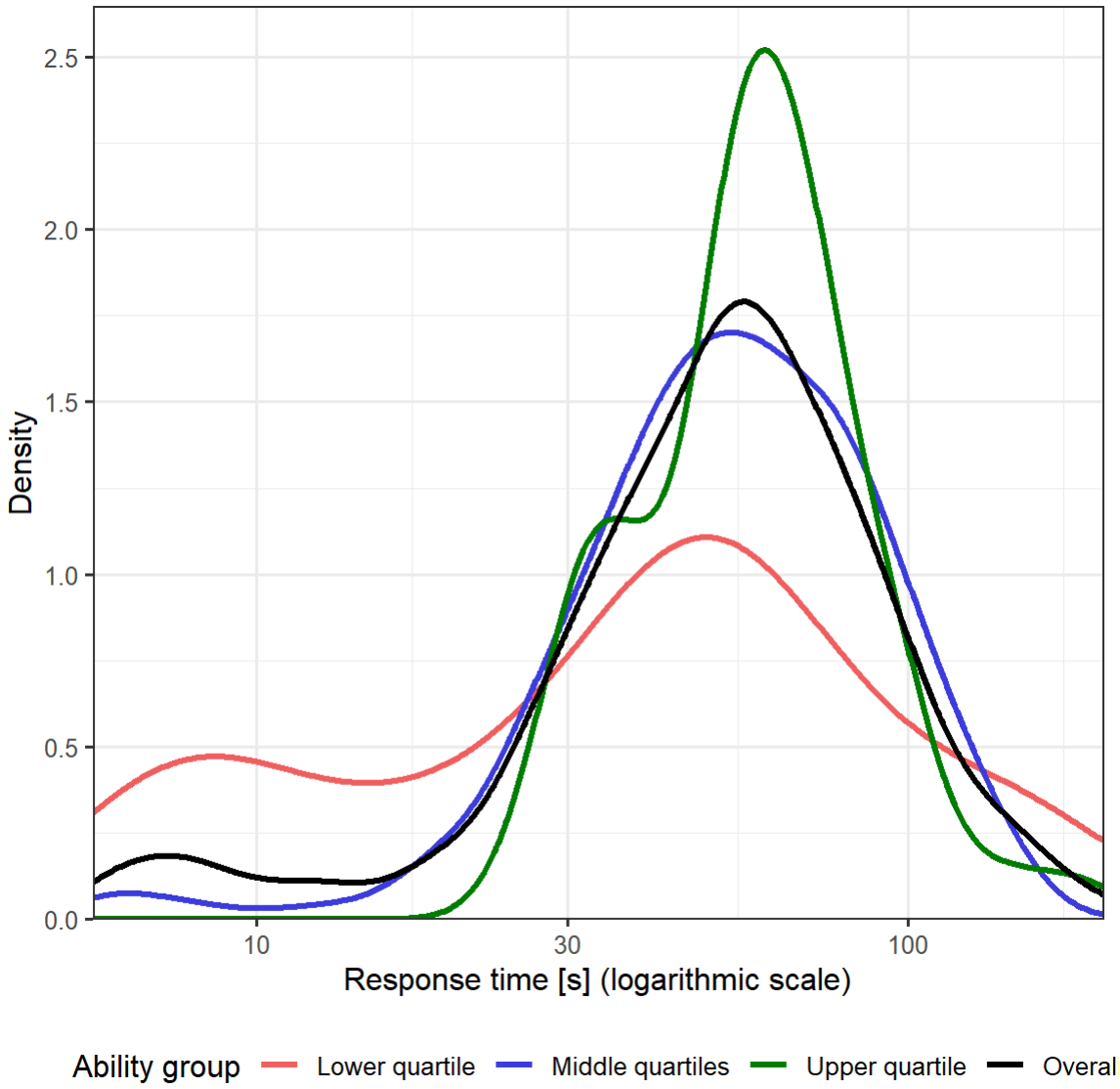
zą gęstość niż woda. Guma ma większą gęstość niż drewno, ale mniejszą niż metal. Met

```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



szą gęstość niż woda. Guma ma większą gęstość niż drewno, ale mniejszą niż metal. Met

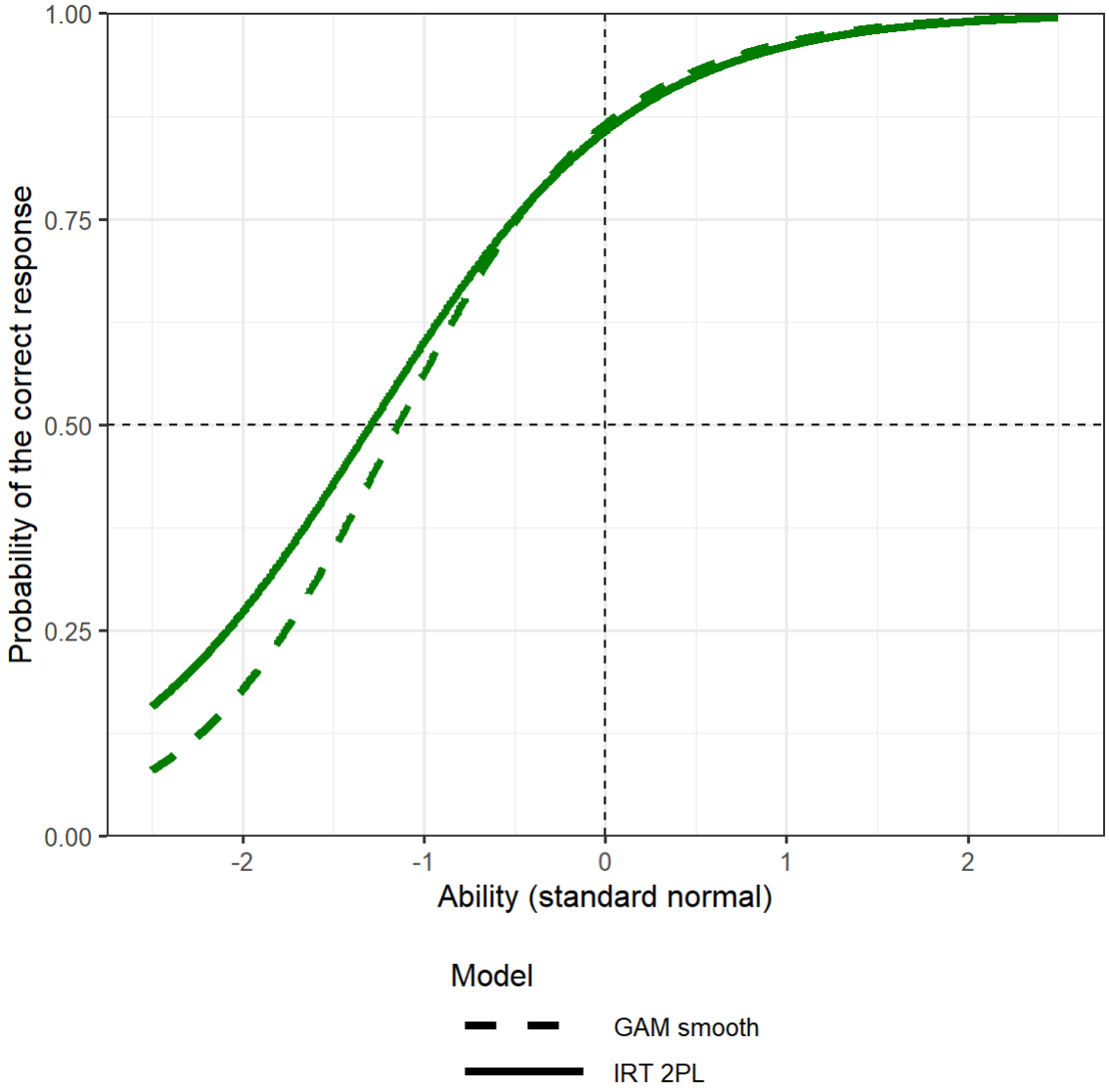
10.2.20.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.6	19.3	41.9	63.3	199.0	3.6	0.9
Middle quartiles	6.0	38.3	53.6	77.5	142.7	4.0	0.5
Upper quartile	24.8	44.7	58.0	72.1	186.9	4.0	0.4
Overall	5.6	36.6	53.2	73.3	199.0	3.9	0.7

10.2.21 S041102 (TIMSS)

10.2.21.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	78.62%	305	TIMSS	MC	0.46	0.46	1.39	−1.29	0.000

10.2.21.2 Gender DIF:

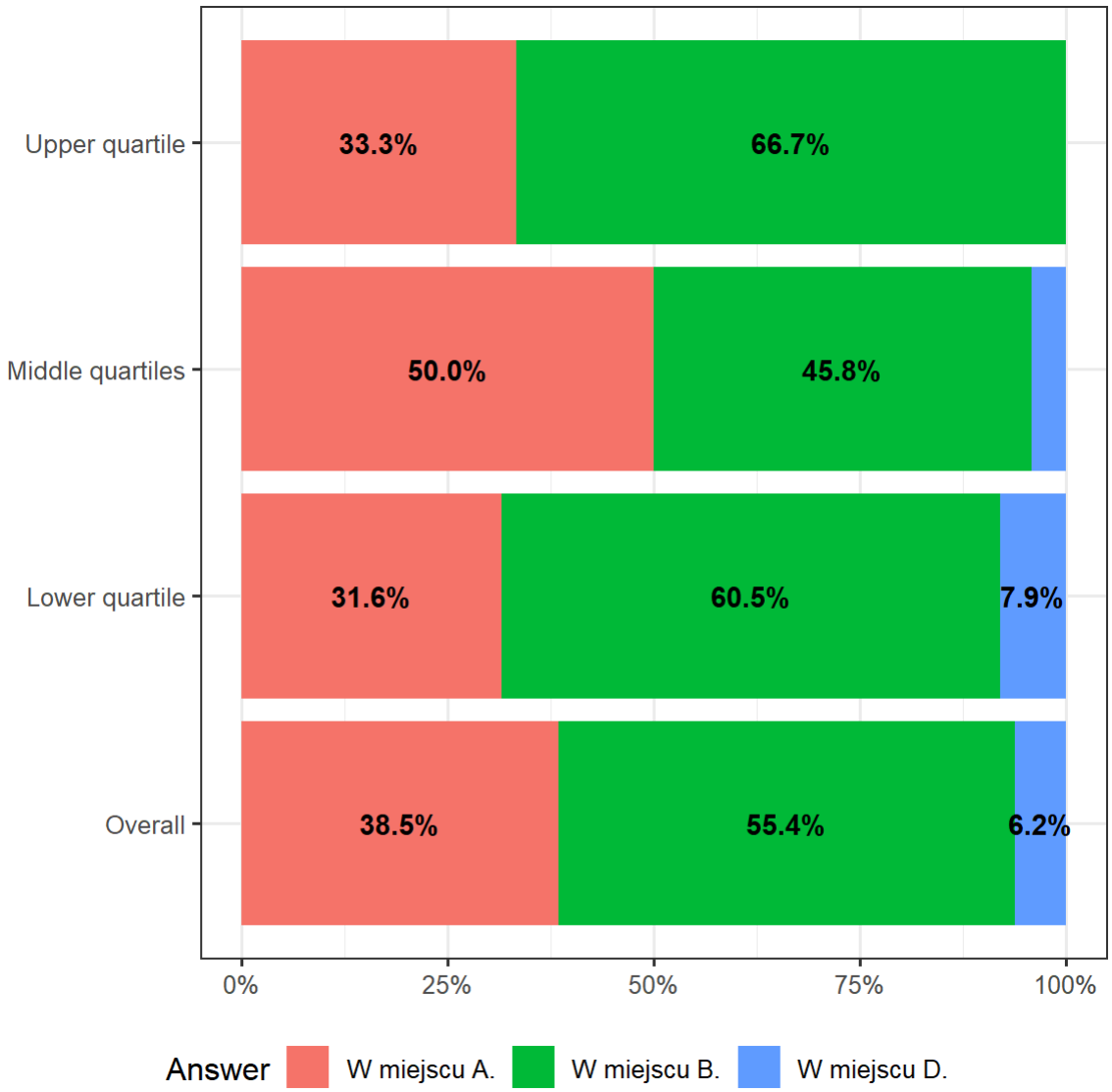
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.29	0.525	1.000	0.005	N/A

10.2.21.3 Frequency of answers:

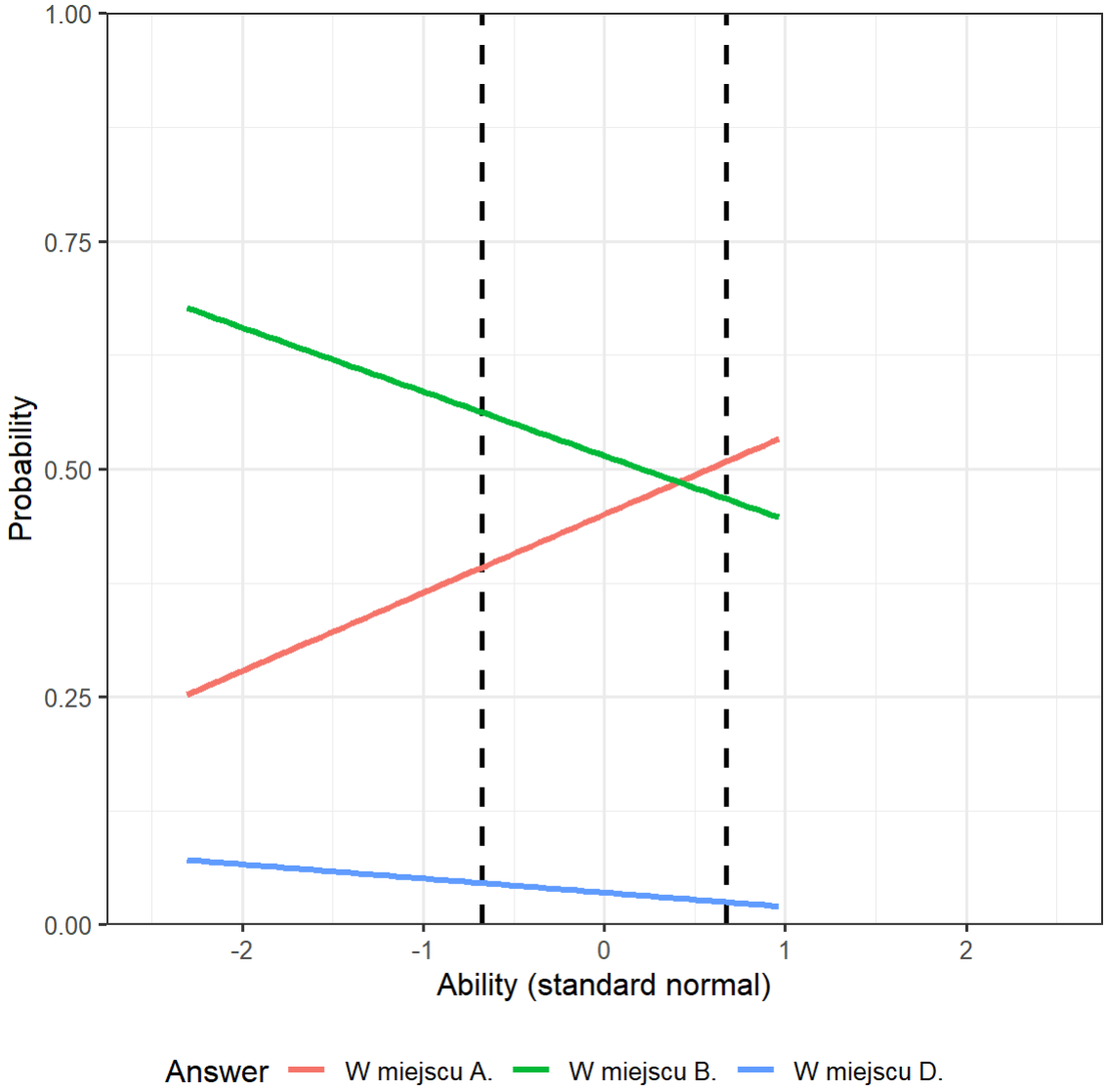
Answer	Frequency	Percent
W miejscu A.	25	8.2%
W miejscu B.	36	11.8%
W miejscu C. (correct)	239	78.6%
W miejscu D.	4	1.3%
Overall	304	100.0%

10.2.21.4 Distractors functioning:

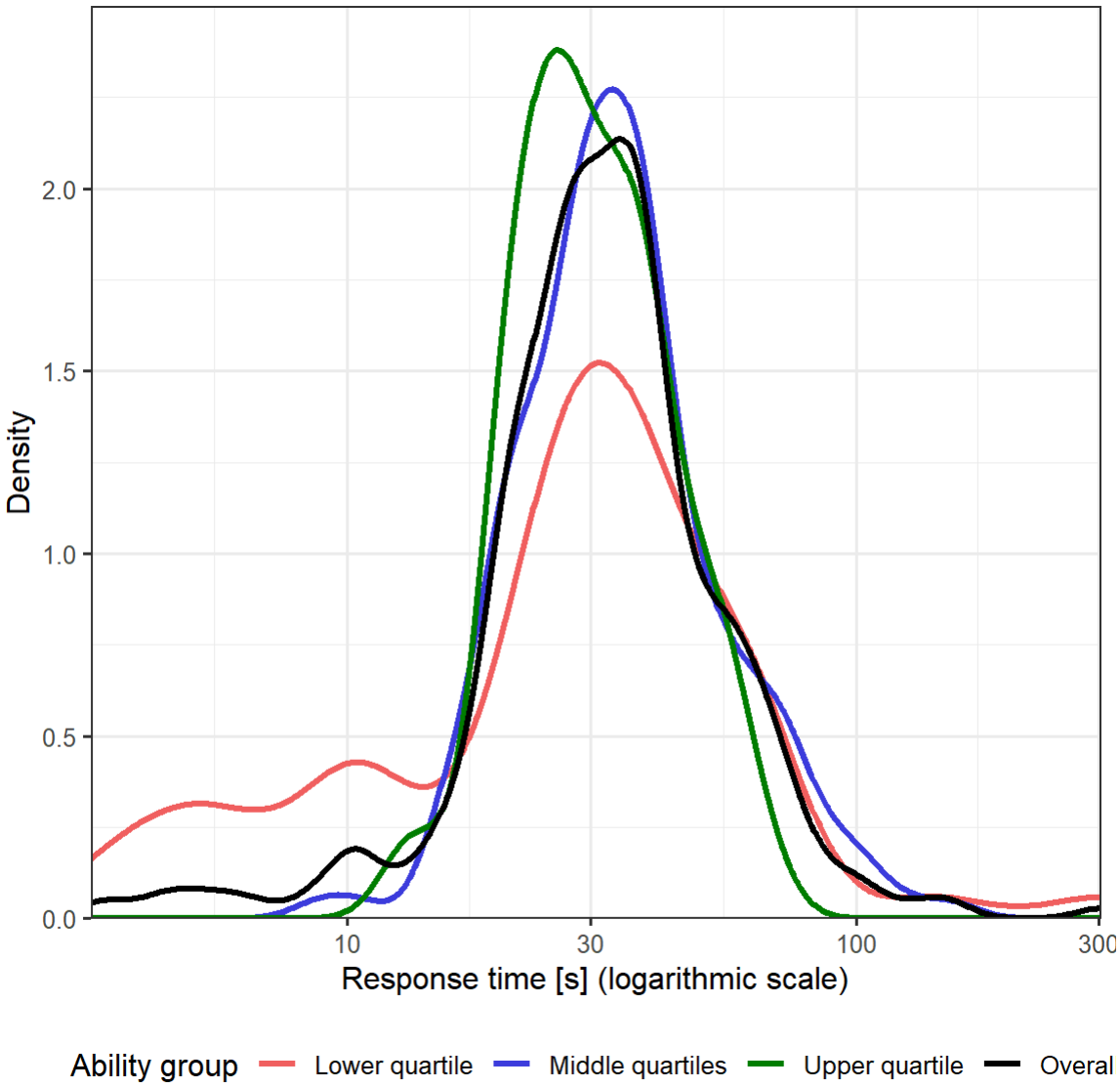
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
W miejscu A.	12	12	1	25	31.6%	50.0%	33.3%	38.5%
W miejscu B.	23	11	2	36	60.5%	45.8%	66.7%	55.4%
W miejscu D.	3	1	0	4	7.9%	4.2%	0.0%	6.2%
Overall	38	24	3	65	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



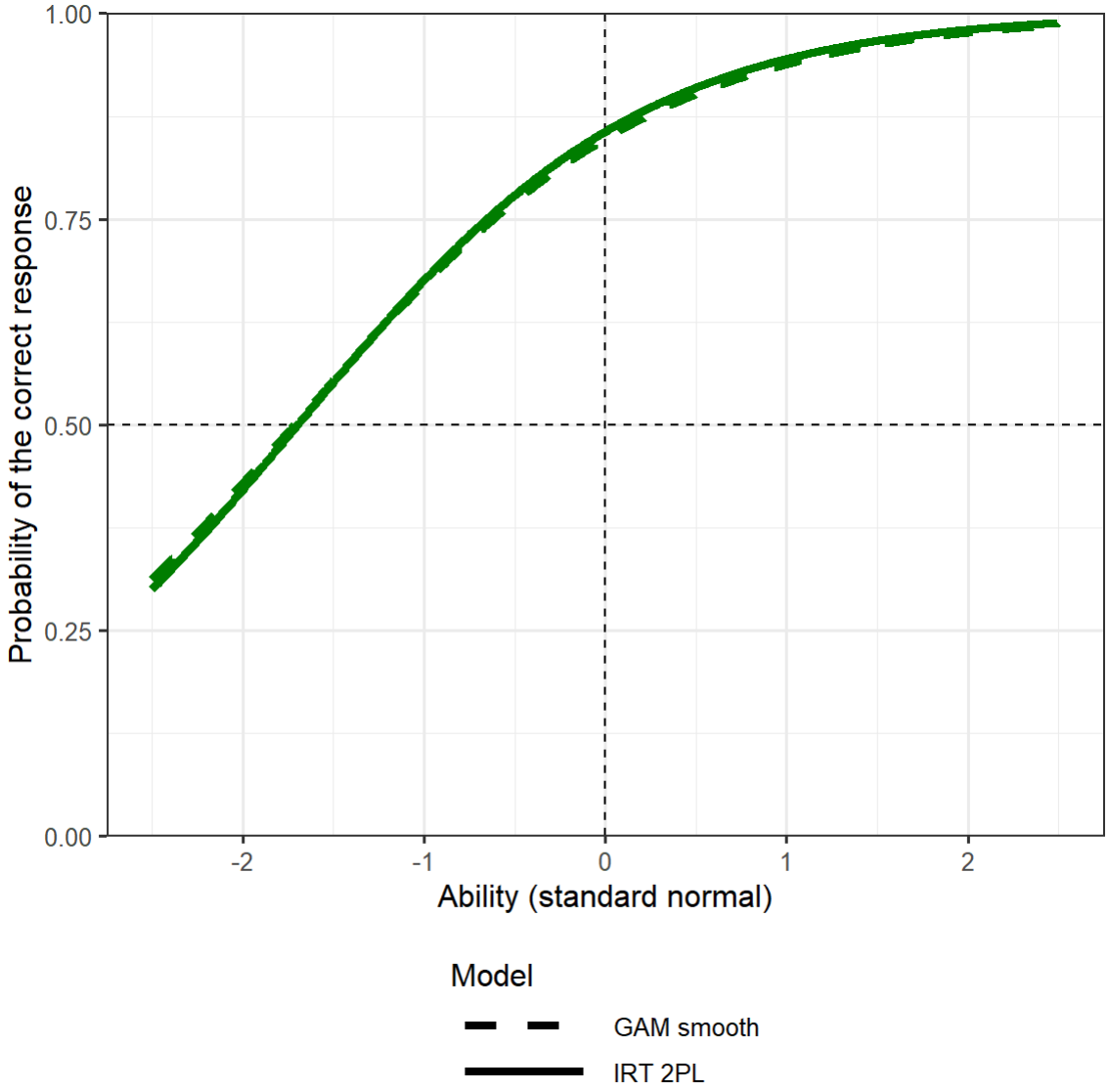
10.2.21.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	17.4	29.0	39.5	301.8	3.2	0.9
Middle quartiles	8.8	25.8	33.3	43.1	153.7	3.5	0.5
Upper quartile	13.3	23.7	29.1	38.5	65.9	3.4	0.4
Overall	3.1	24.4	31.9	40.9	301.8	3.4	0.6

10.2.22 S061028 (TIMSS)

10.2.22.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	80.92%	305	TIMSS	MC	0.33	0.38	1.05	−1.69	0.020

10.2.22.2 Gender DIF:

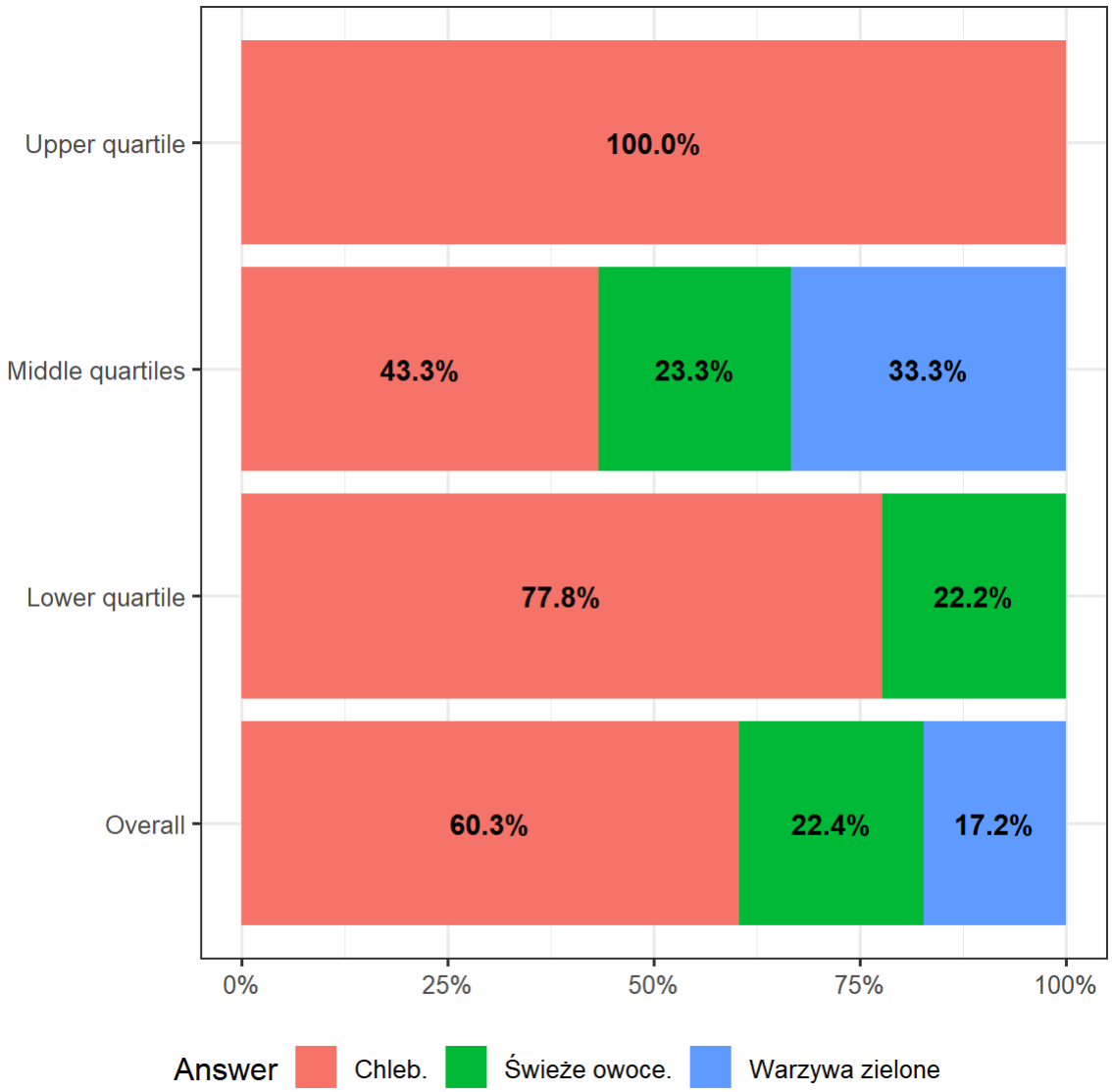
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.43	0.297	1.000	0.009	N/A

10.2.22.3 Frequency of answers:

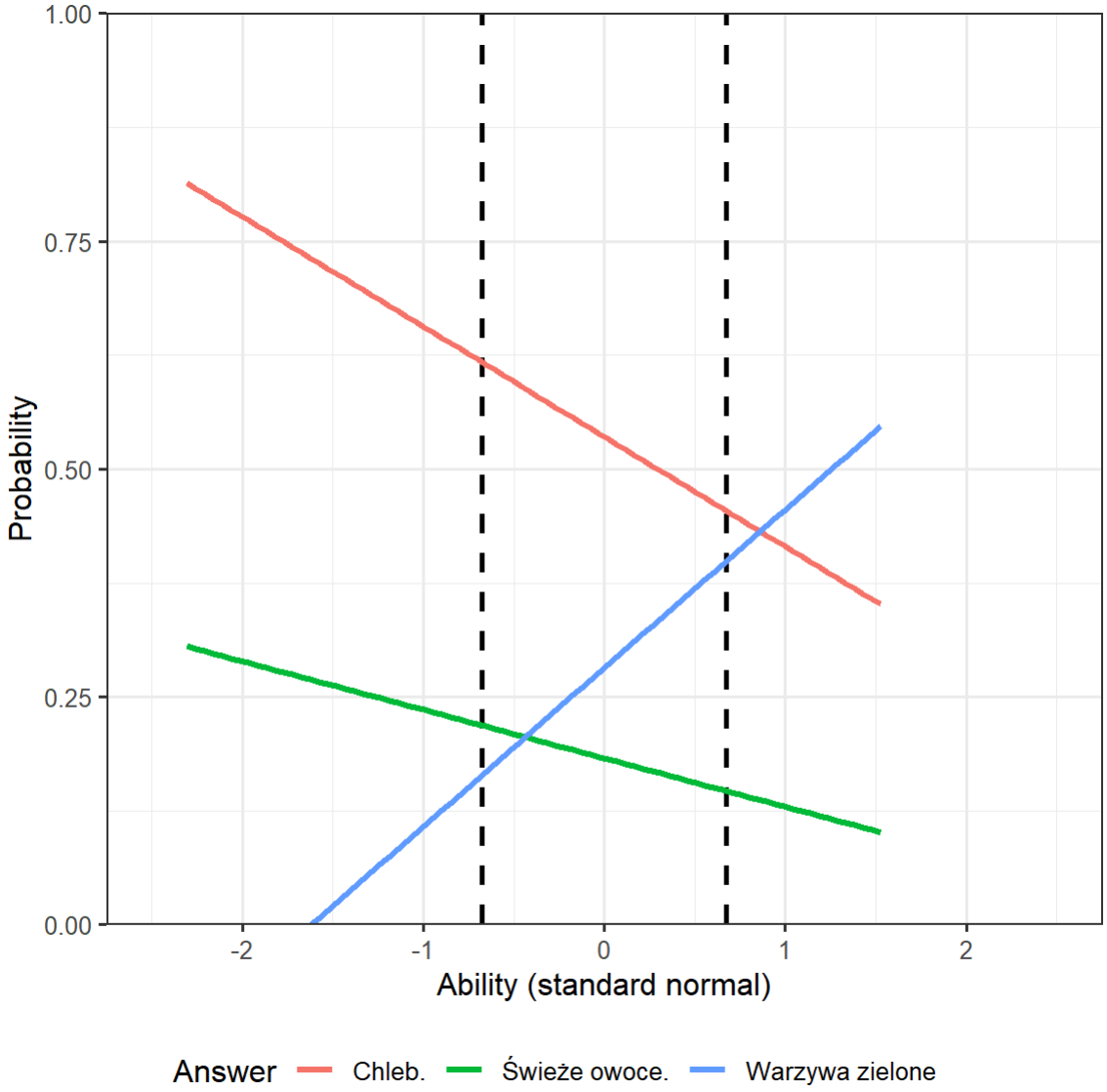
Answer	Frequency	Percent
Chleb.	35	11.5%
Ryba. (correct)	246	80.9%
Warzywa zielone	10	3.3%
Świeże owoce.	13	4.3%
Overall	304	100.0%

10.2.22.4 Distractors functioning:

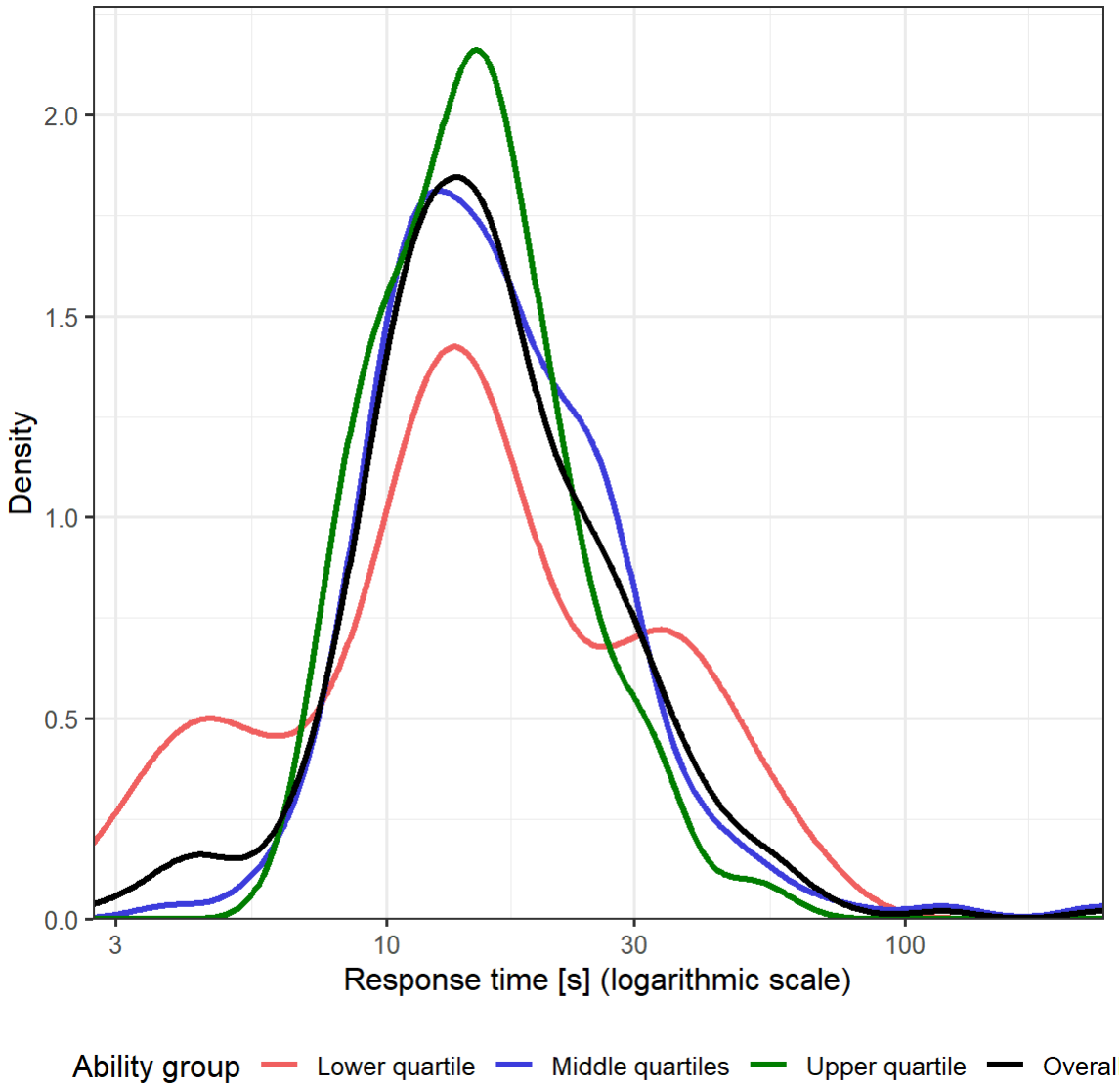
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Chleb.	21	13	1	35	77.8%	43.3%	100.0%	60.3%
Świeże owoce.	6	7	0	13	22.2%	23.3%	0.0%	22.4%
Overall	27	30	1	58	100.0%	100.0%	100.0%	100.0%
Warzywa zielone	0	10	0	10	0.0%	33.3%	0.0%	17.2%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



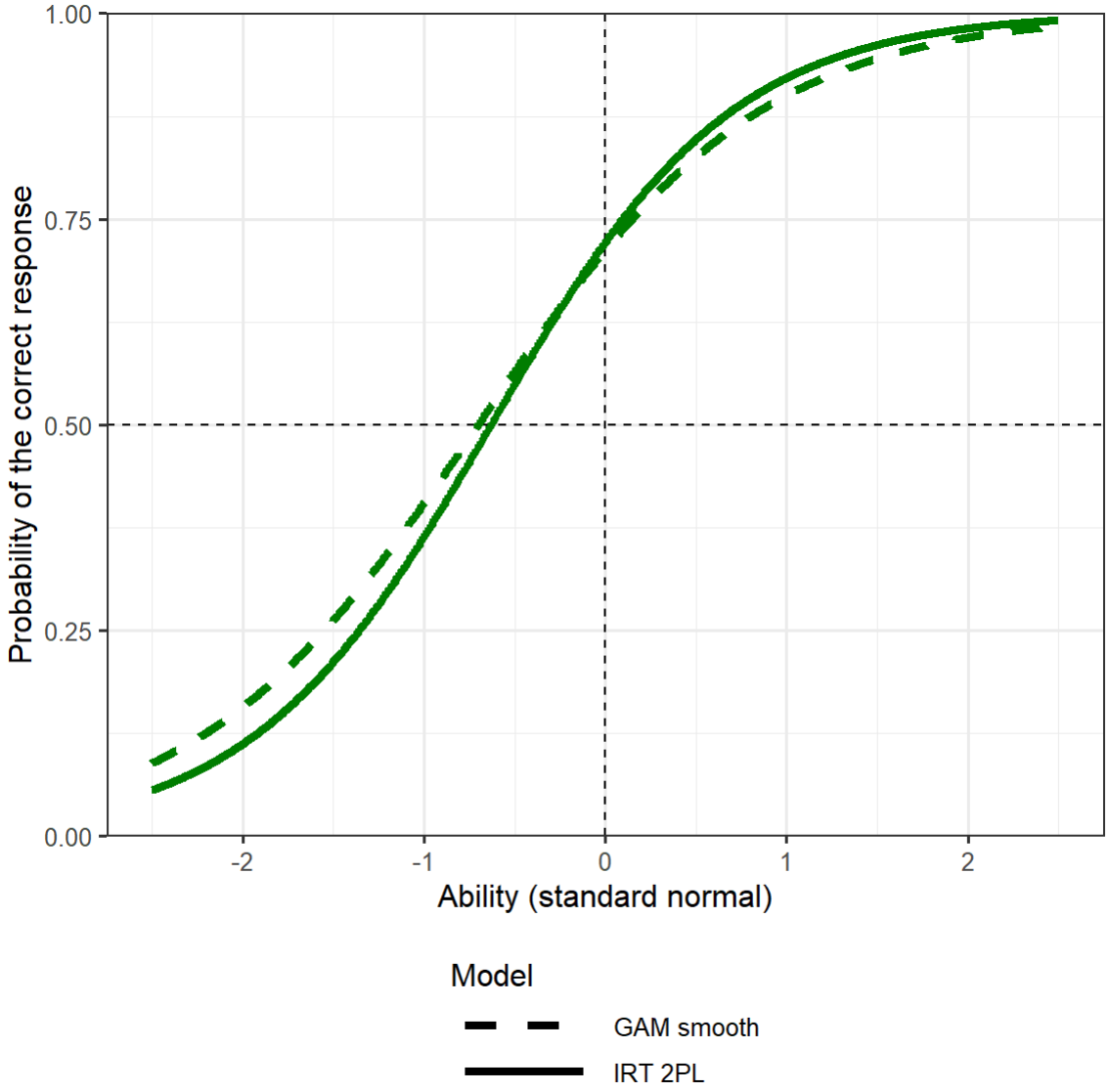
10.2.22.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	10.1	13.6	24.1	61.5	2.7	0.8
Middle quartiles	3.8	11.3	15.2	22.6	241.1	2.8	0.5
Upper quartile	7.3	11.0	14.6	19.3	51.4	2.7	0.4
Overall	2.7	10.8	14.8	21.9	241.1	2.7	0.6

10.2.23 S090064 (TIMSS)

10.2.23.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	66.07%	281	LLM teachers' selection	MC	0.48	0.34	1.51	−0.63	0.0(

10.2.23.2 Gender DIF:

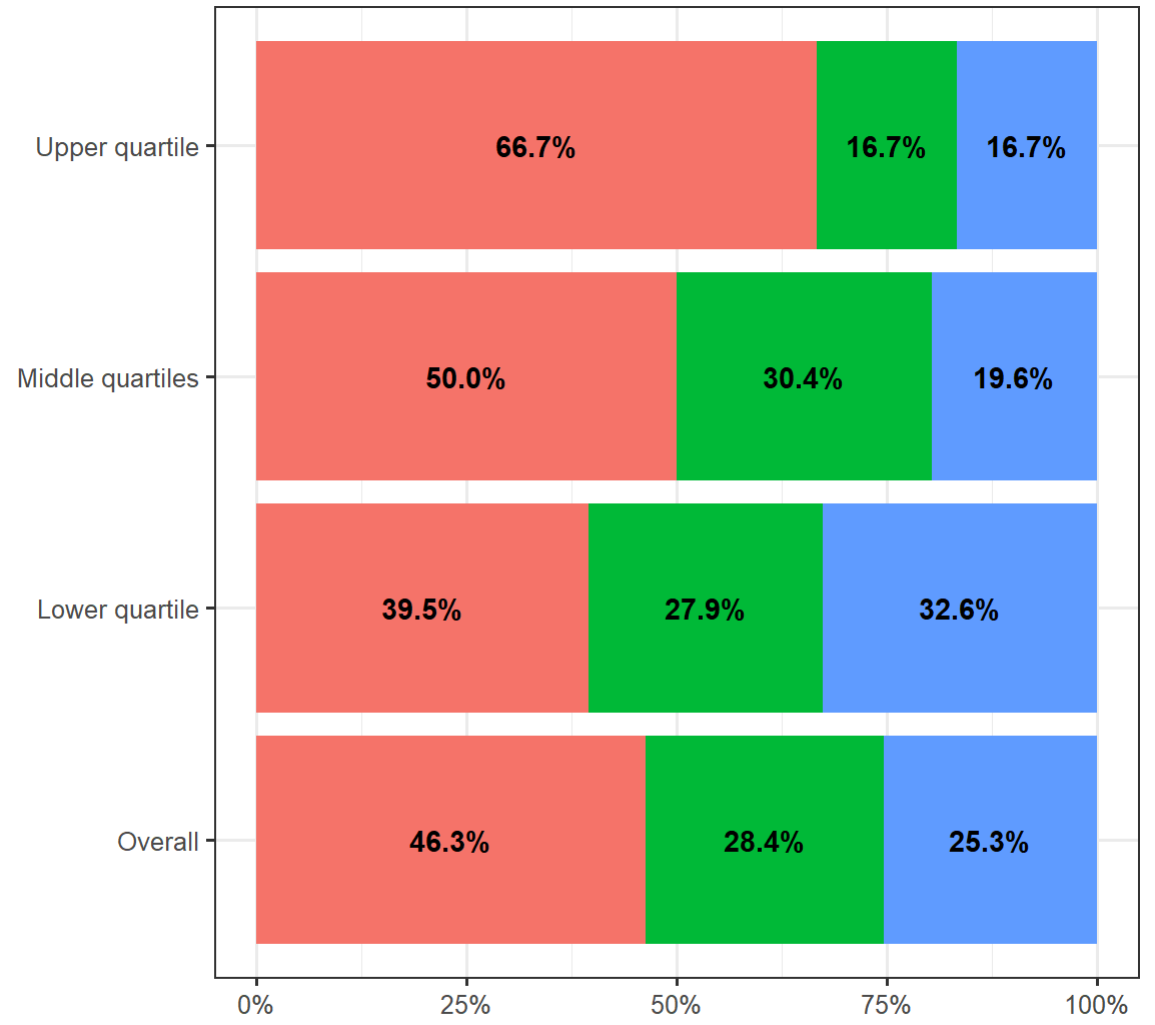
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.47	0.480	1.000	0.005	N/A

10.2.23.3 Frequency of answers:

Answer	Frequency	Percent
Woda zmienia się z ciała stałego w ciecz.	44	15.7%
Woda zmienia się z cieczy w ciało stałe. (correct)	185	66.1%
Woda zmienia się z cieczy w gaz.	27	9.6%
Woda zmienia się z gazu w ciecz.	24	8.6%
Overall	280	100.0%

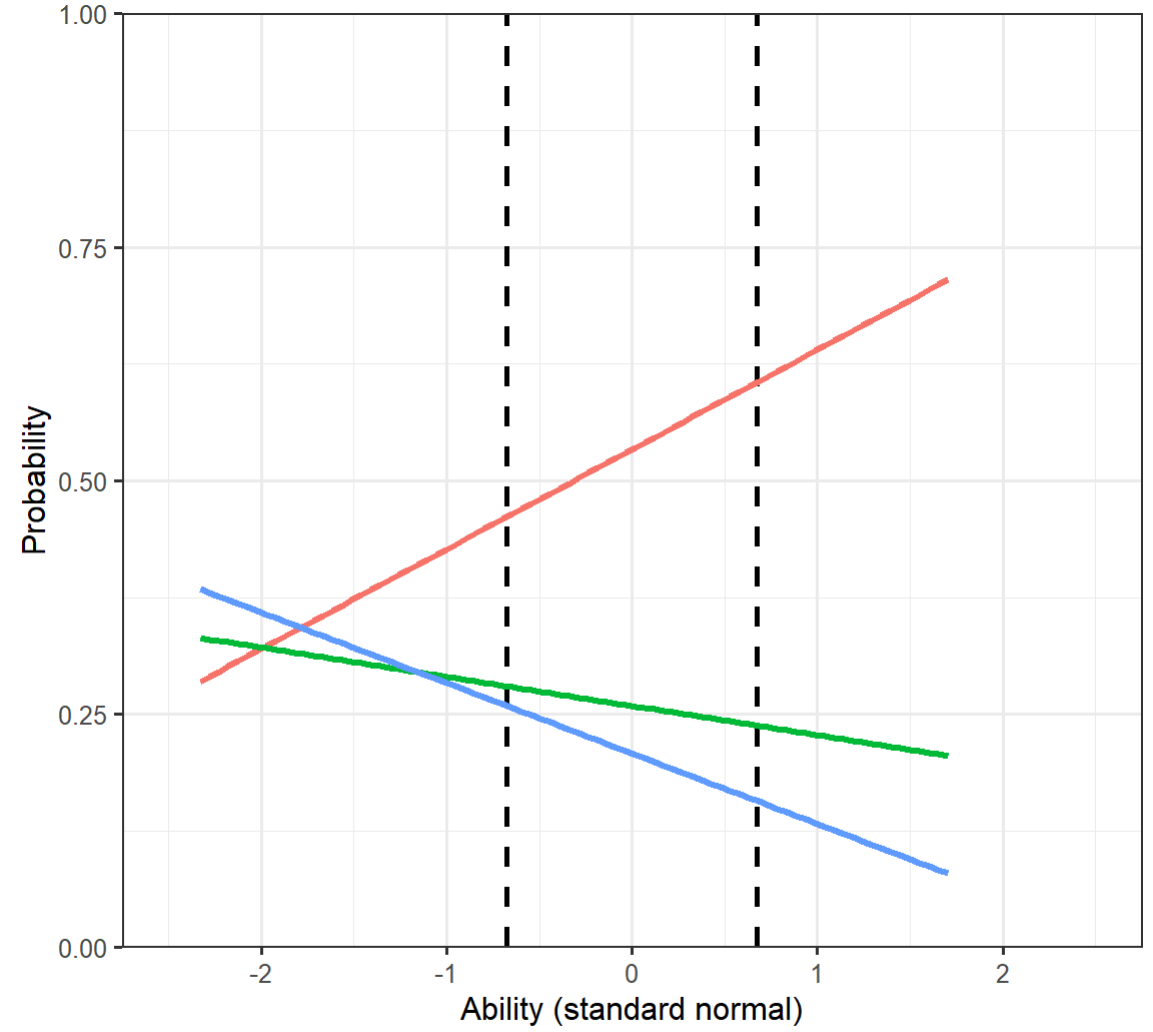
10.2.23.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Woda zmienia się z ciała stałego w ciecz.	17	23	4	44	39.5%	50.0%	66.7%	46.3%
Woda zmienia się z cieczy w gaz.	12	14	1	27	27.9%	30.4%	16.7%	28.4%
Woda zmienia się z gazu w ciecz.	14	9	1	24	32.6%	19.6%	16.7%	25.3%
Overall	43	46	6	95	100.0%	100.0%	100.0%	100.0%



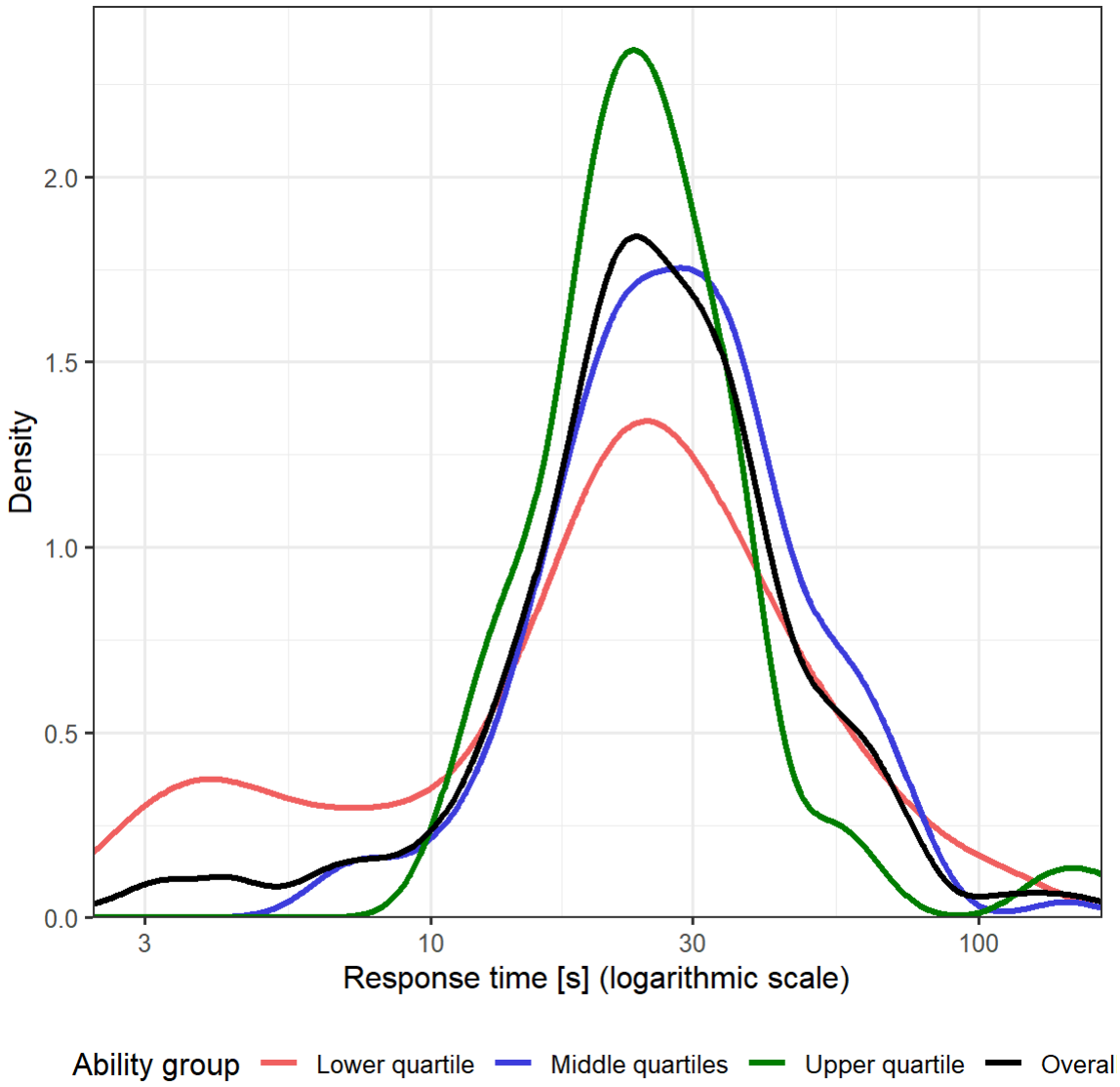
ɔda zmienia się z ciała stałego w ciecz. Woda zmienia się z cieczy w gaz. Woda zmienia

```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



ɔda zmienia się z ciała stałego w ciecz. Woda zmienia się z cieczy w gaz. Woda zmienia

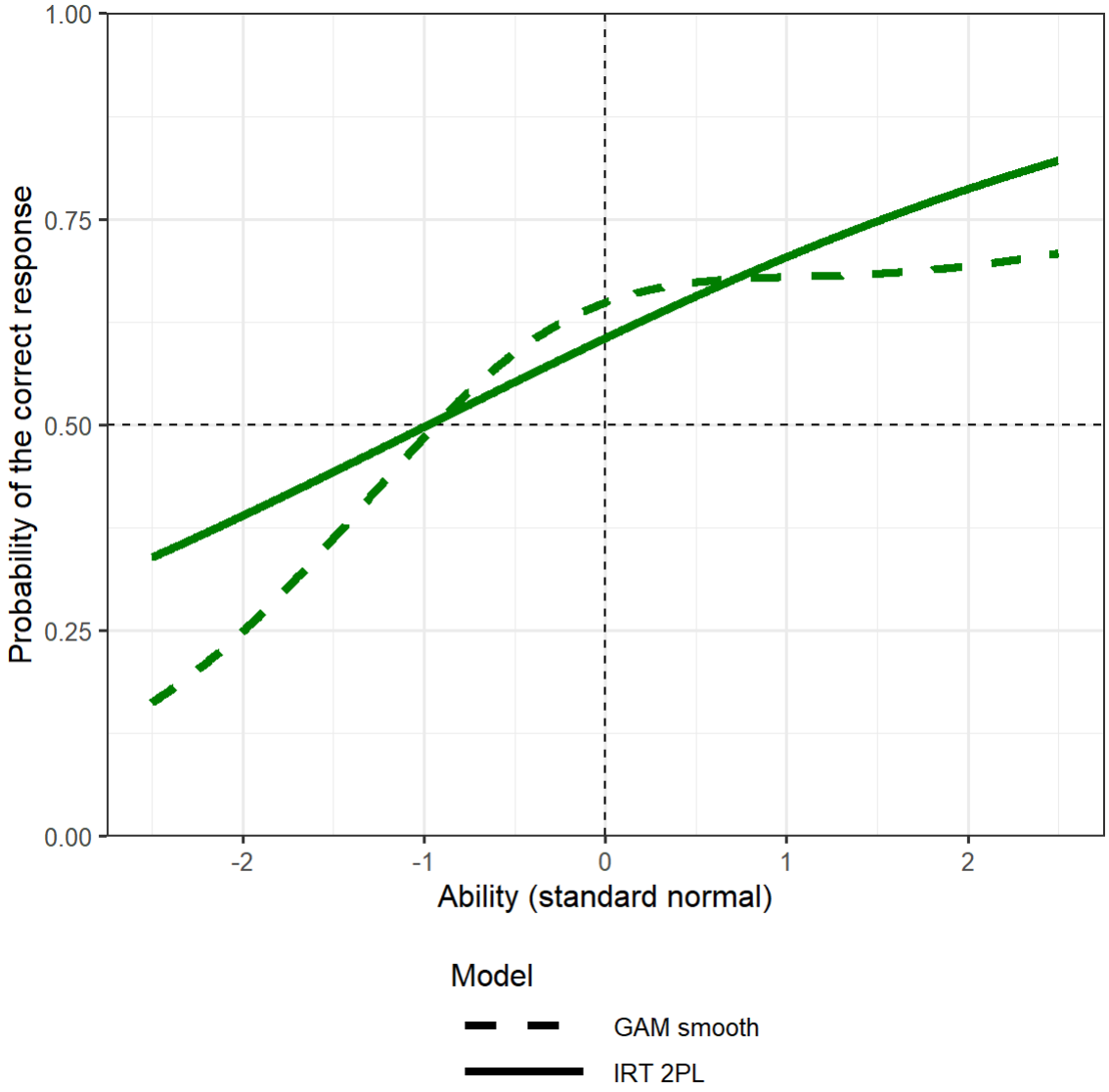
10.2.23.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.4	13.8	22.9	34.2	111.3	3.0	0.9
Middle quartiles	6.4	20.1	28.0	37.6	144.6	3.3	0.5
Upper quartile	10.3	18.8	24.2	30.8	167.1	3.2	0.5
Overall	2.4	18.5	25.1	35.5	167.1	3.2	0.6

10.2.24 S090051 (TIMSS)

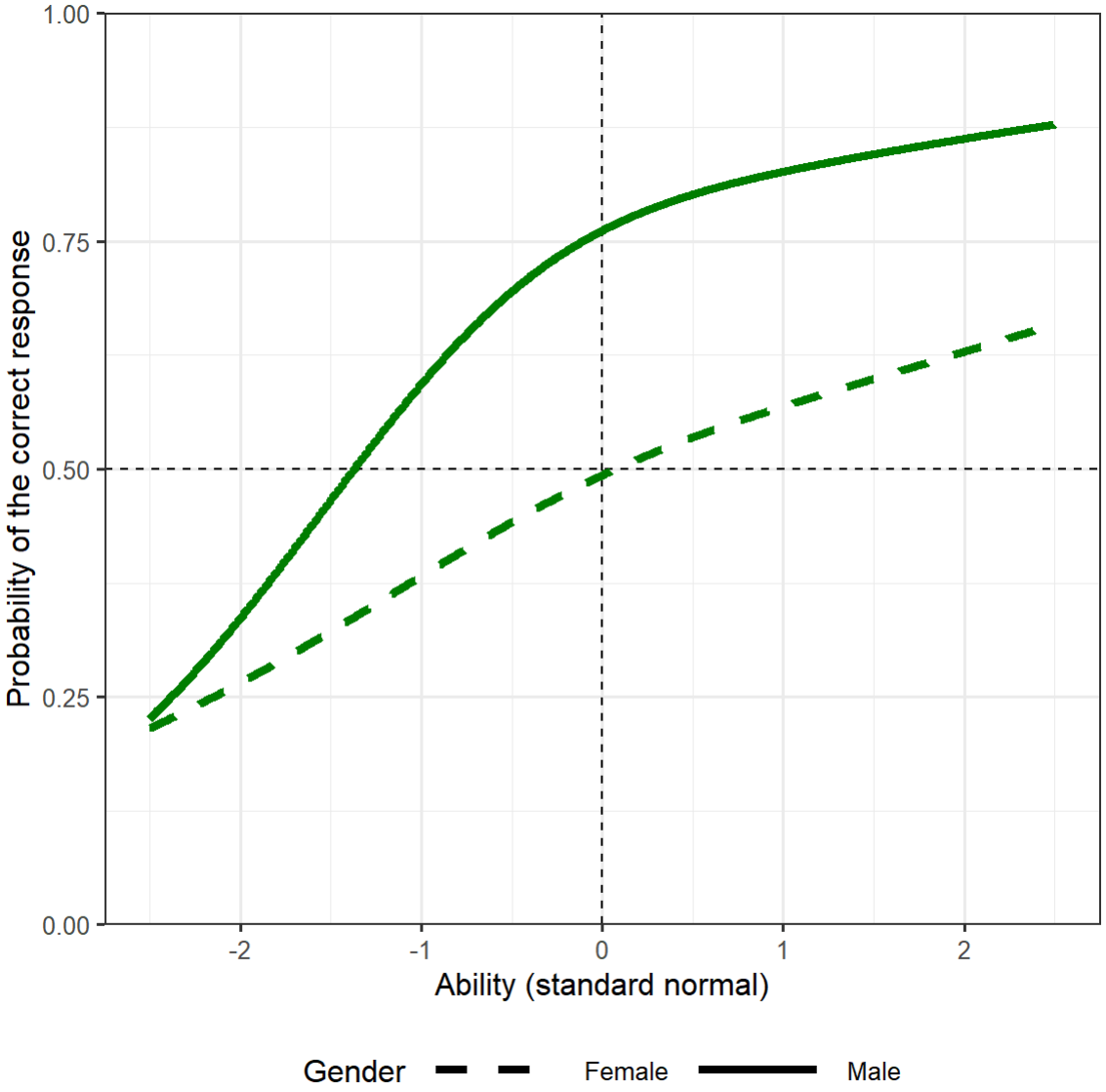
10.2.24.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	59.87%	304	LLM teachers' selection	MC	0.20	0.09	0.44	−0.97	0.02

10.2.24.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
17.99	0.000	0.012	0.051	moderate



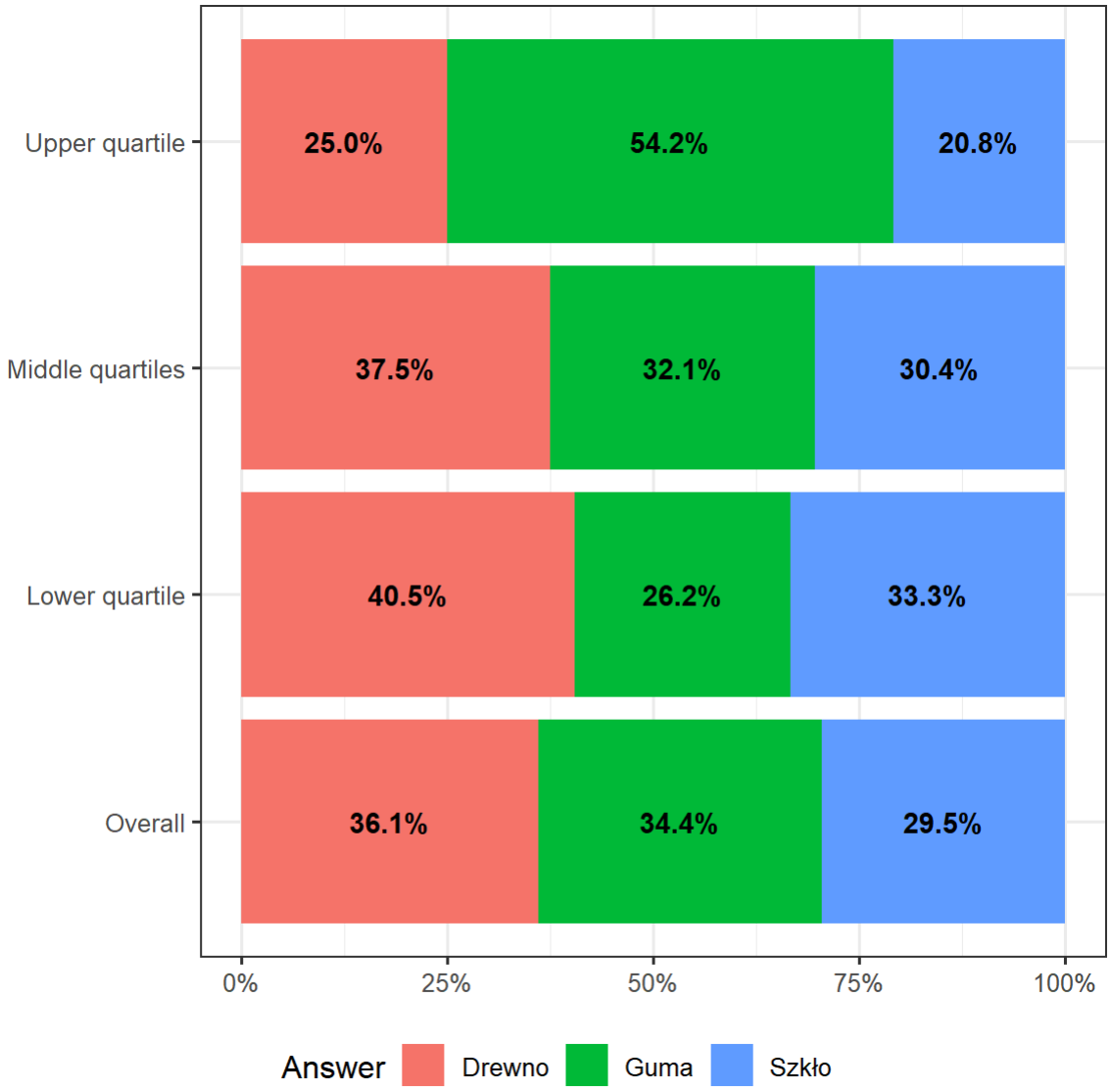
10.2.24.3 Frequency of answers:

Answer	Frequency	Percent
Drewno	44	14.5%
Guma	42	13.8%
Miedź (correct)	182	59.9%
Szkło	36	11.8%
Overall	304	100.0%

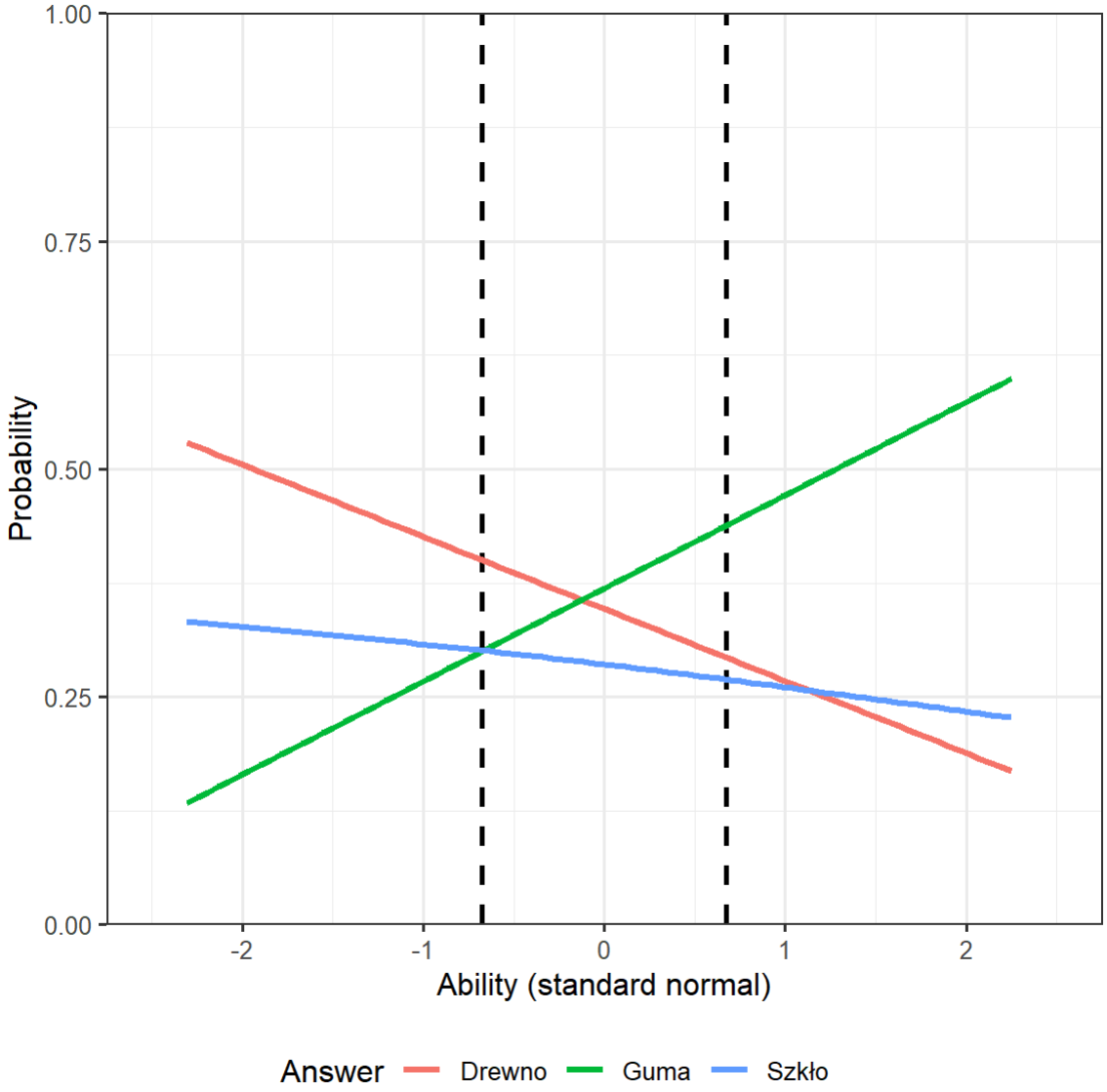
10.2.24.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Drewno	17	21	6	44	40.5%	37.5%	25.0%	36.1%

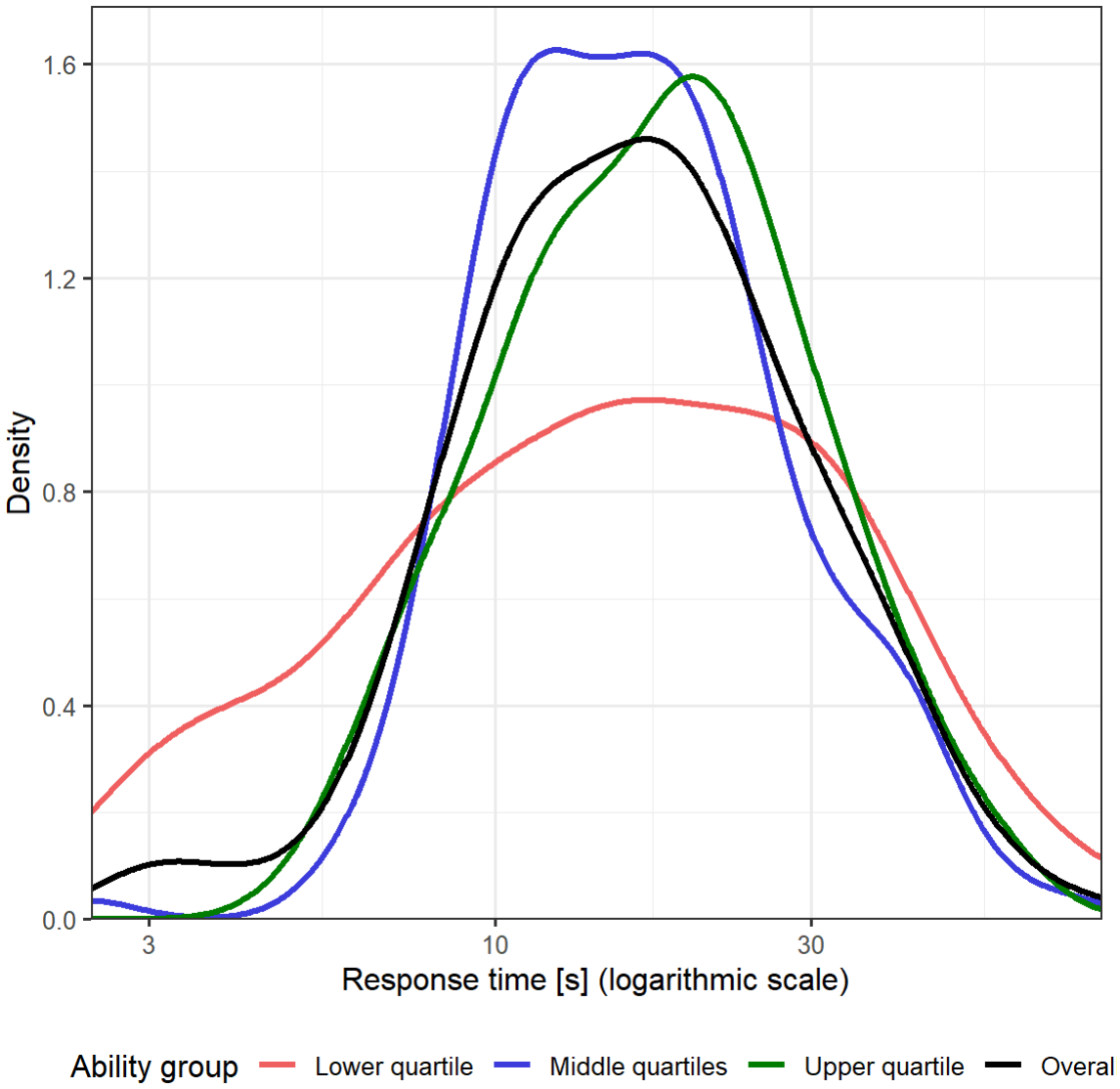
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Guma	11	18	13	42	26.2%	32.1%	54.2%	34.4%
Szkło	14	17	5	36	33.3%	30.4%	20.8%	29.5%
Overall	42	56	24	122	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



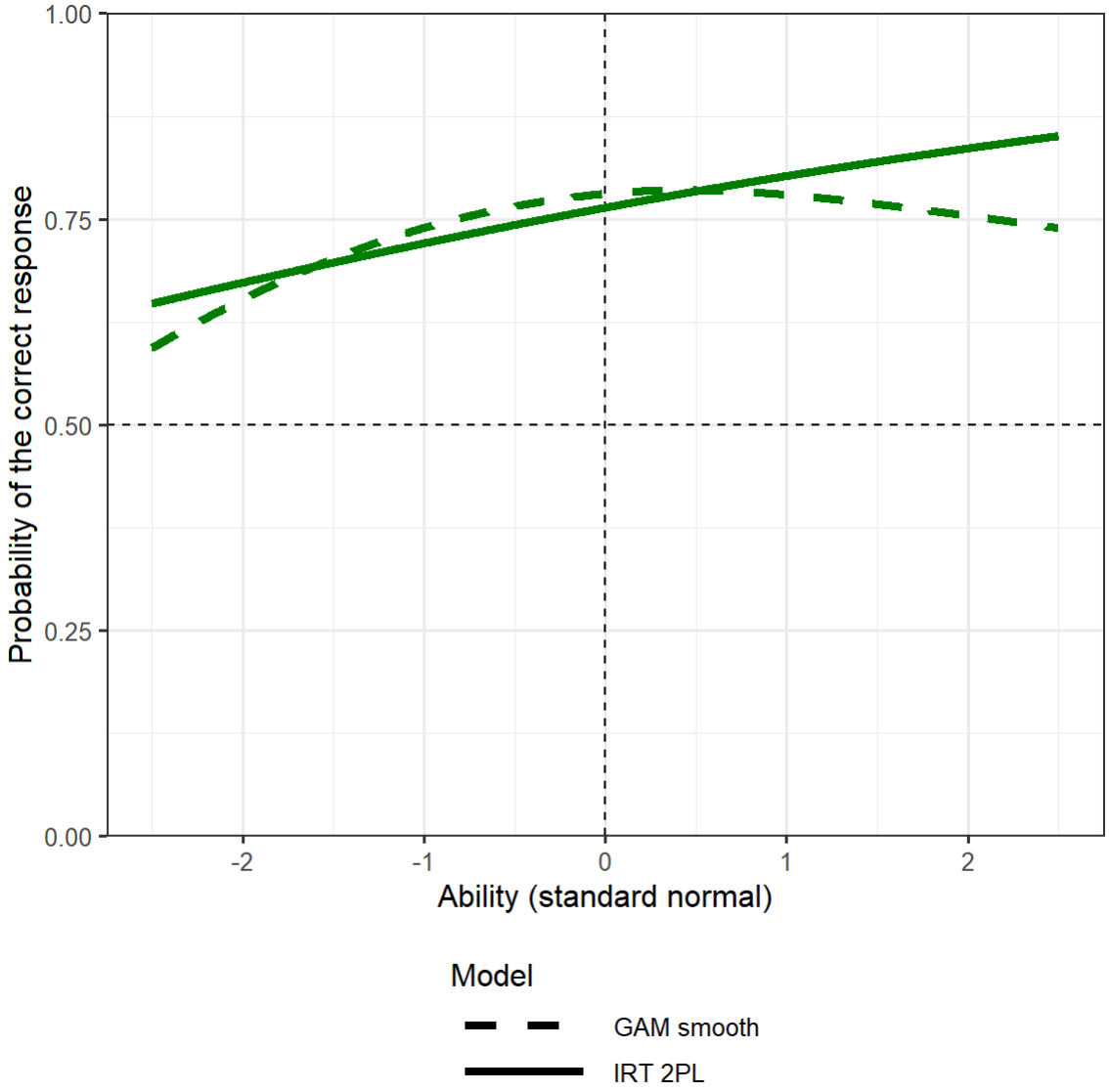
10.2.24.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.8	8.0	15.8	27.5	82.4	2.7	0.8
Middle quartiles	2.4	11.1	15.7	21.9	73.3	2.8	0.5
Upper quartile	5.6	11.9	18.1	24.7	57.2	2.9	0.5
Overall	2.4	10.7	16.1	24.4	82.4	2.8	0.6

10.2.25 S890021 (TIMSS)

10.2.25.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	76.18%	319	LLM complex setup	MC	0.06	0.00	0.23	−5.18	0.01

10.2.25.2 Gender DIF:

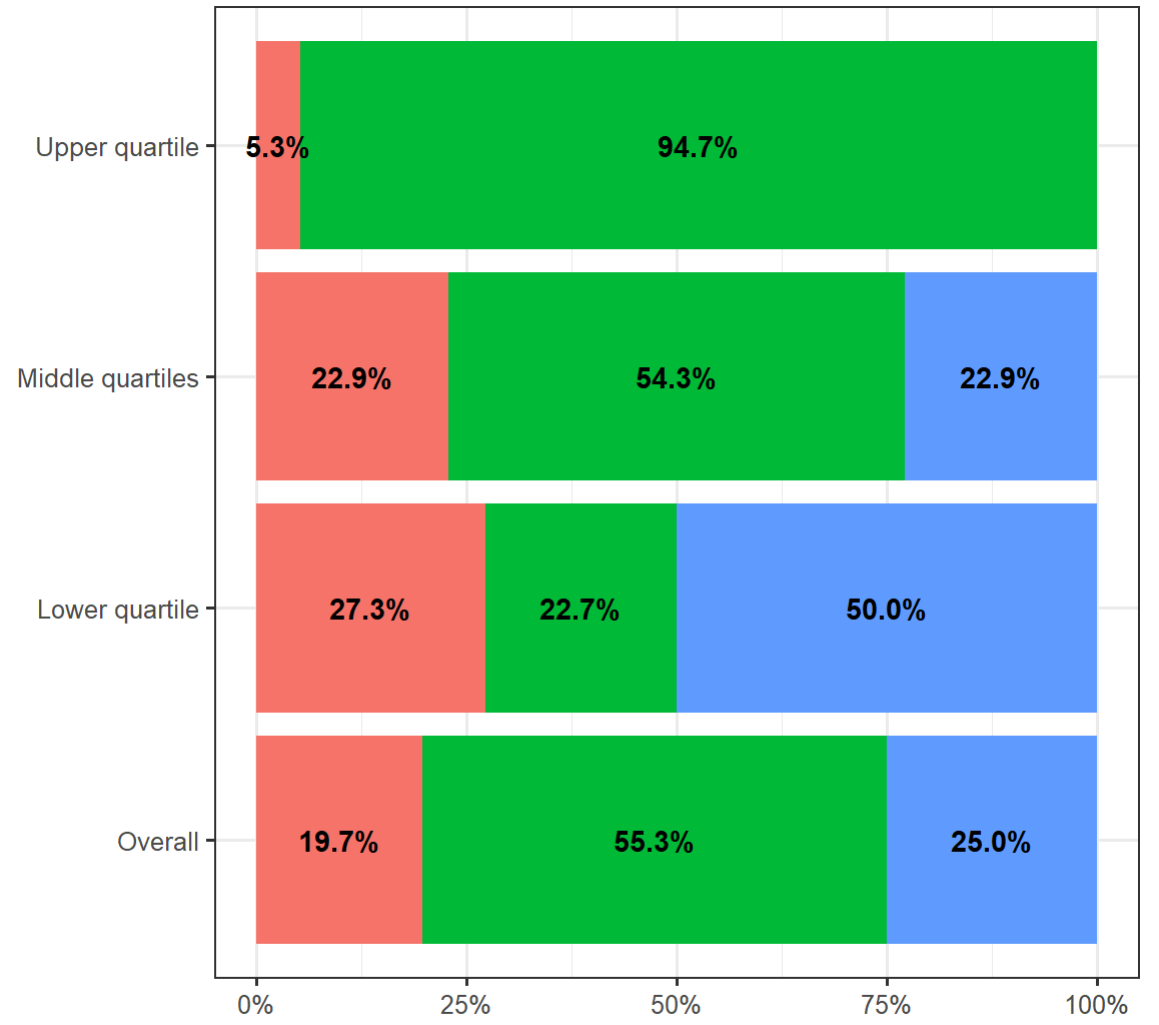
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.53	0.171	1.000	0.012	N/A

10.2.25.3 Frequency of answers:

Answer	Frequency	Percent
Czajnik wchłania wodę z otoczenia.	15	4.7%
Para wodna w czajniku zamienia się w płyn po schłodzeniu.	42	13.2%
Para wodna w czajniku zamieniła się w lód.	19	6.0%
Para wodna z czajnika skrapla się na jego powierzchni. (correct)	243	76.2%
Overall	319	100.0%

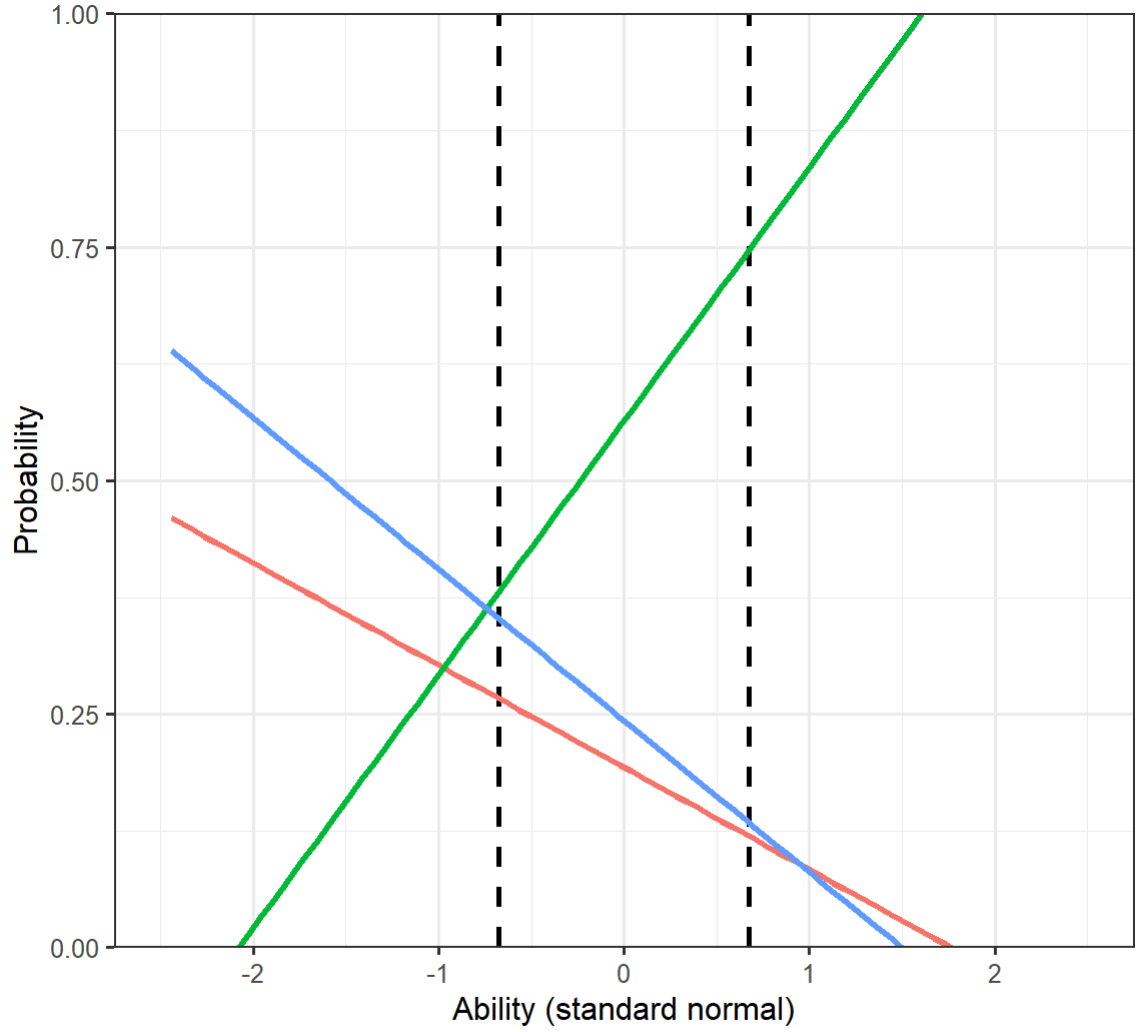
10.2.25.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Czajnik wchłania wodę z otoczenia.	6	8	1	15	27.3%	22.9%	5.3%	19.7%
Para wodna w czajniku zamienia się w płyn po schłodzeniu.	5	19	18	42	22.7%	54.3%	94.7%	55.3%
Para wodna w czajniku zamieniła się w lód.	11	8	0	19	50.0%	22.9%	0.0%	25.0%
Overall	22	35	19	76	100.0%	100.0%	100.0%	100.0%



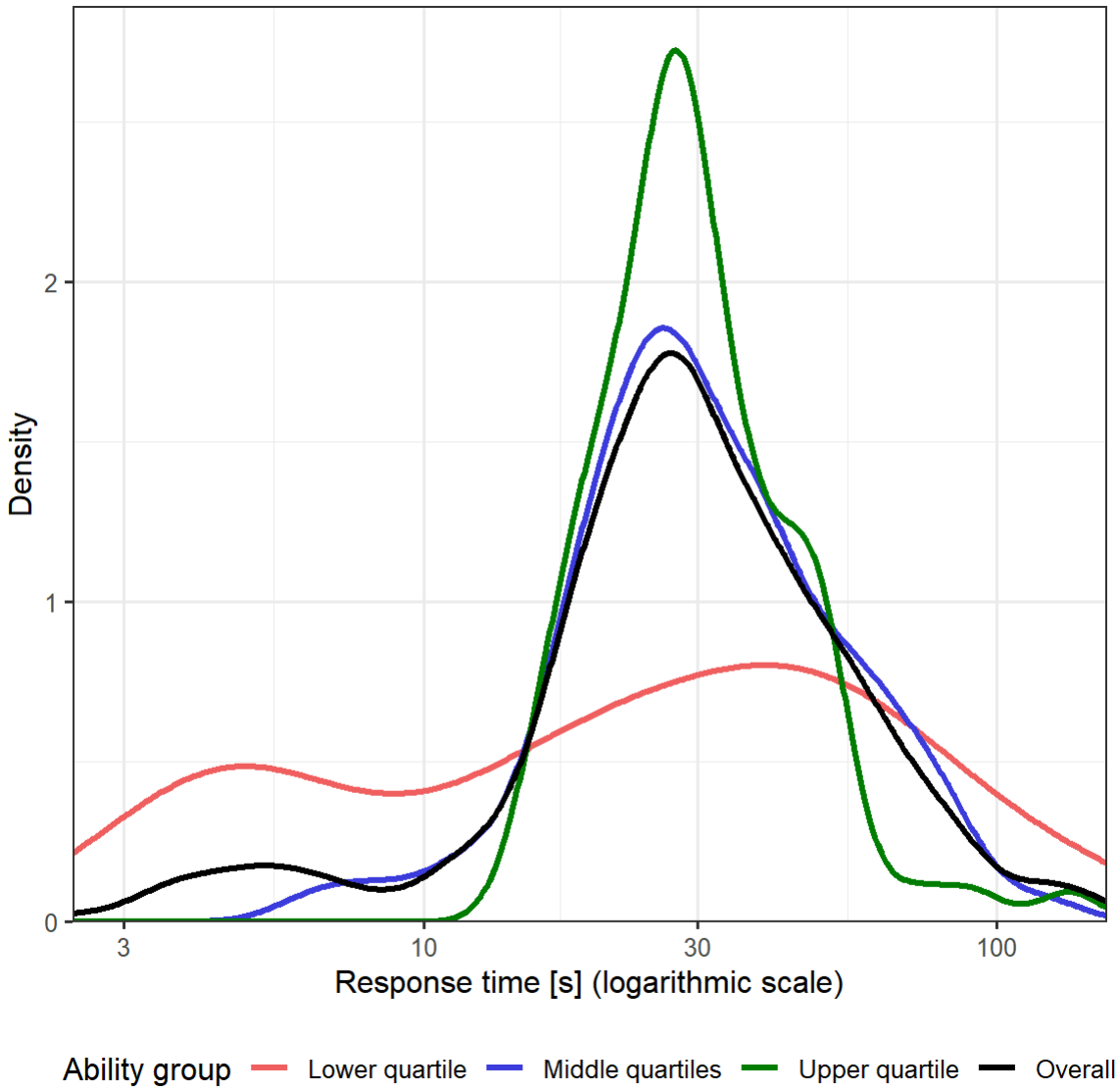
wodę z otoczenia. Para wodna w czajniku zamienia się w płyn po schłodzeniu. Para wodı

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



wodę z otoczenia. Para wodna w czajniku zamienia się w płyn po schłodzeniu. Para wodı

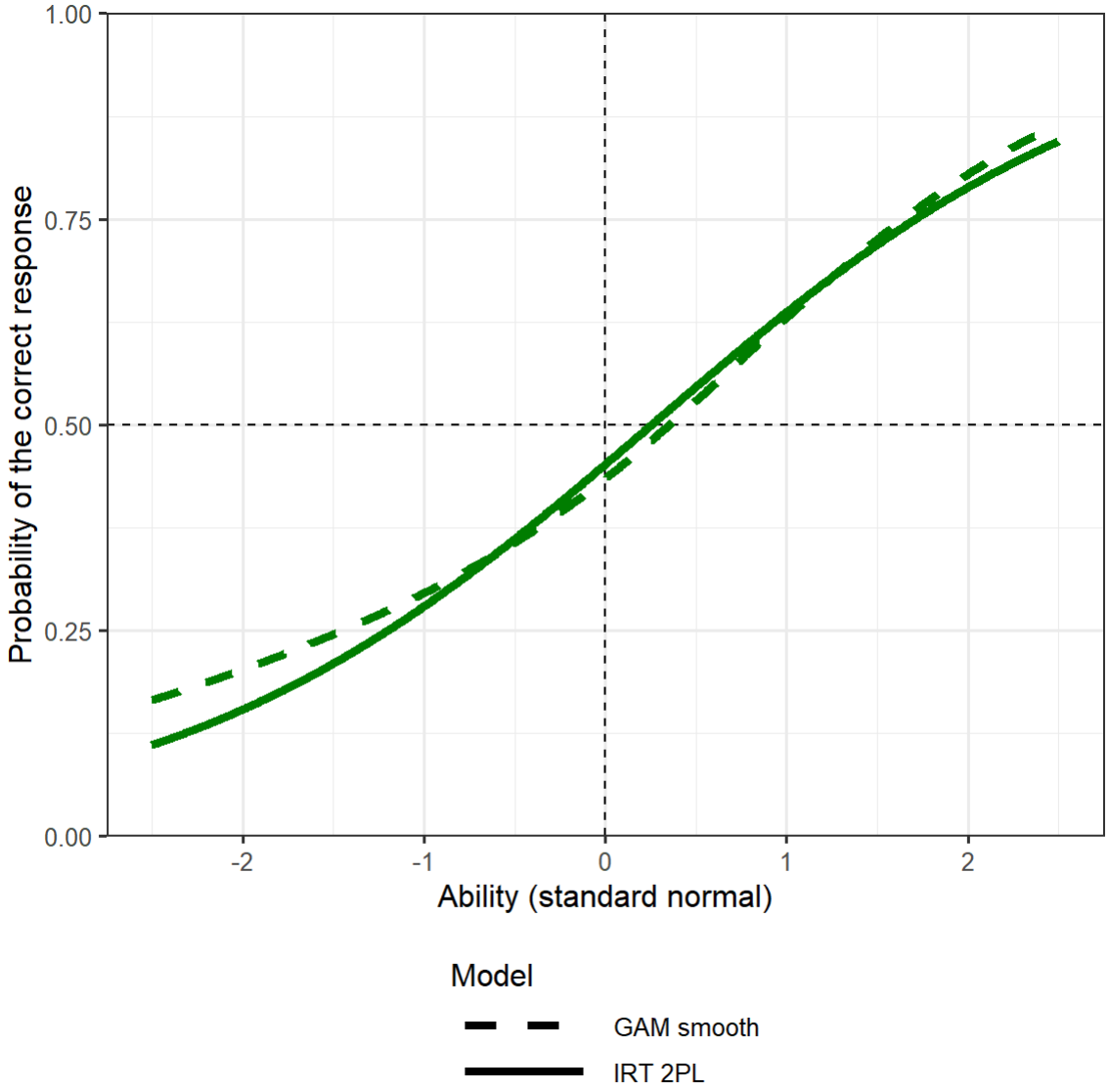
10.2.25.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.4	8.6	25.4	49.9	155.0	3.1	1.1
Middle quartiles	6.2	21.6	28.6	41.9	123.2	3.4	0.5
Upper quartile	14.9	23.3	28.5	36.4	133.8	3.4	0.4
Overall	2.4	20.7	28.5	42.0	155.0	3.3	0.7

10.2.26 S090015 (LLM)

10.2.26.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	45.39%	304	LLM teachers' selection	MC	0.30	0.16	0.75	0.26	0.02

10.2.26.2 Gender DIF:

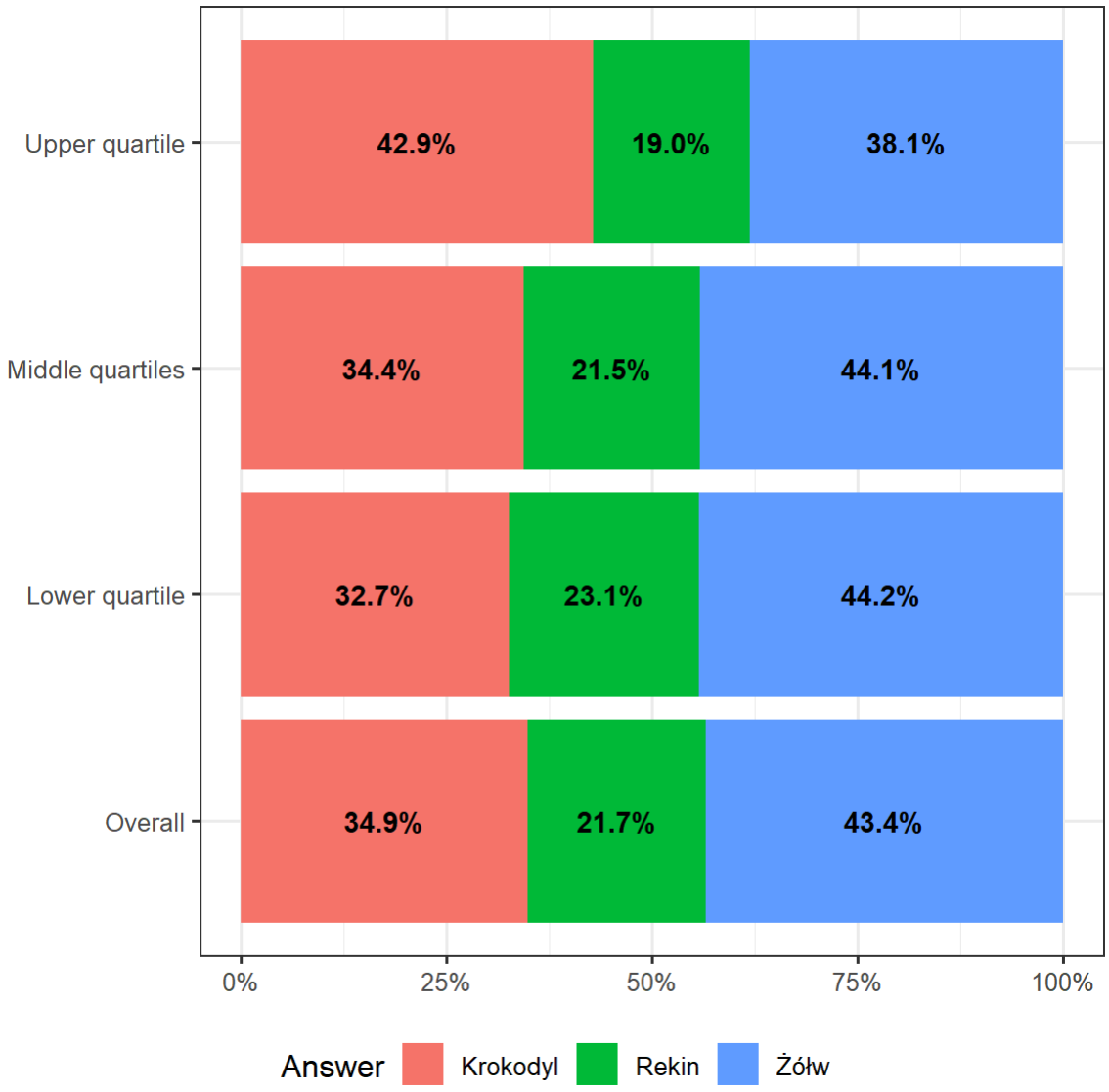
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.12	0.941	1.000	0.000	N/A

10.2.26.3 Frequency of answers:

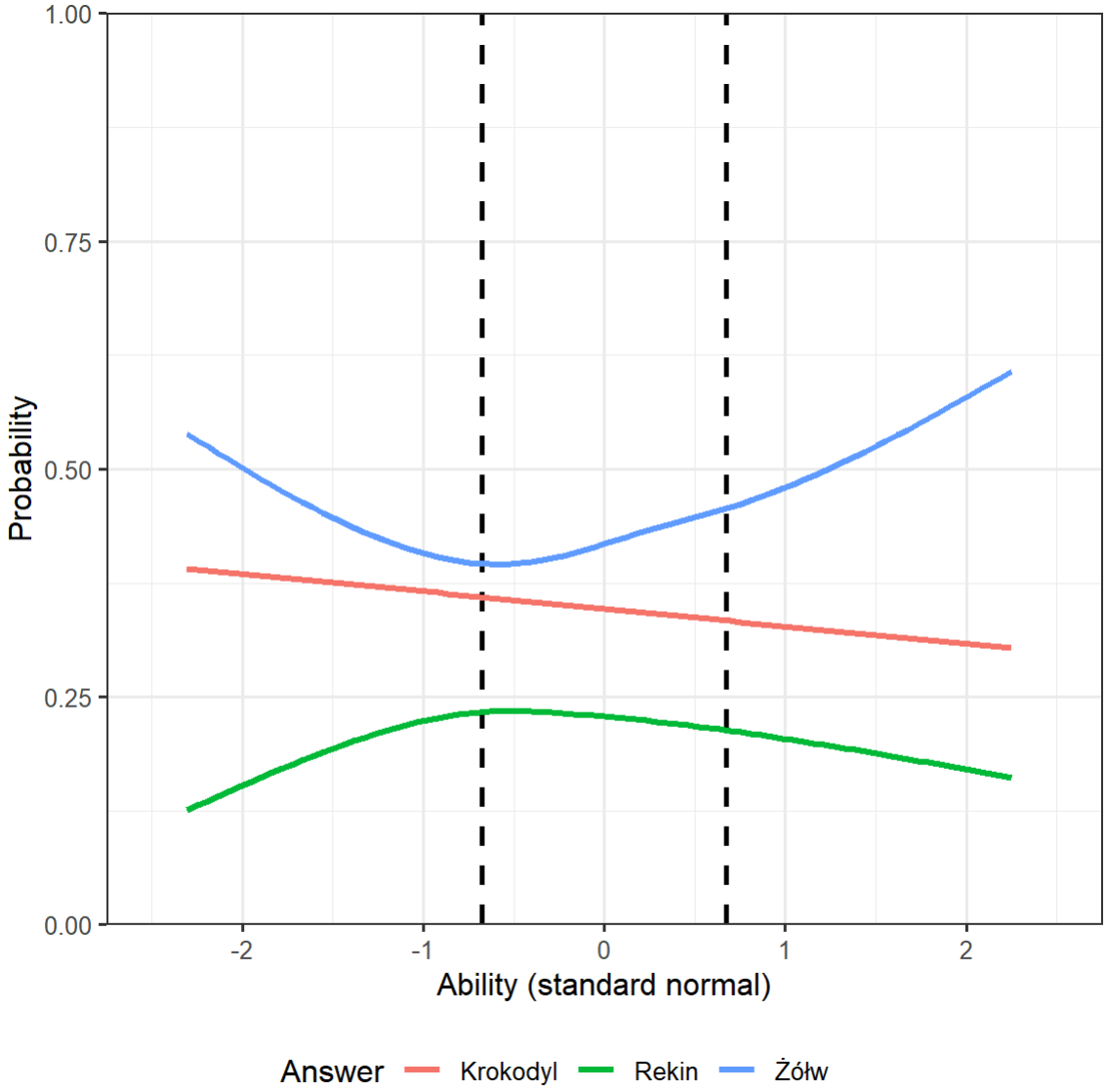
Answer	Frequency	Percent
Delfin (correct)	138	45.4%
Krokodyl	58	19.1%
Rekin	36	11.8%
Żółw	72	23.7%
Overall	304	100.0%

10.2.26.4 Distractors functioning:

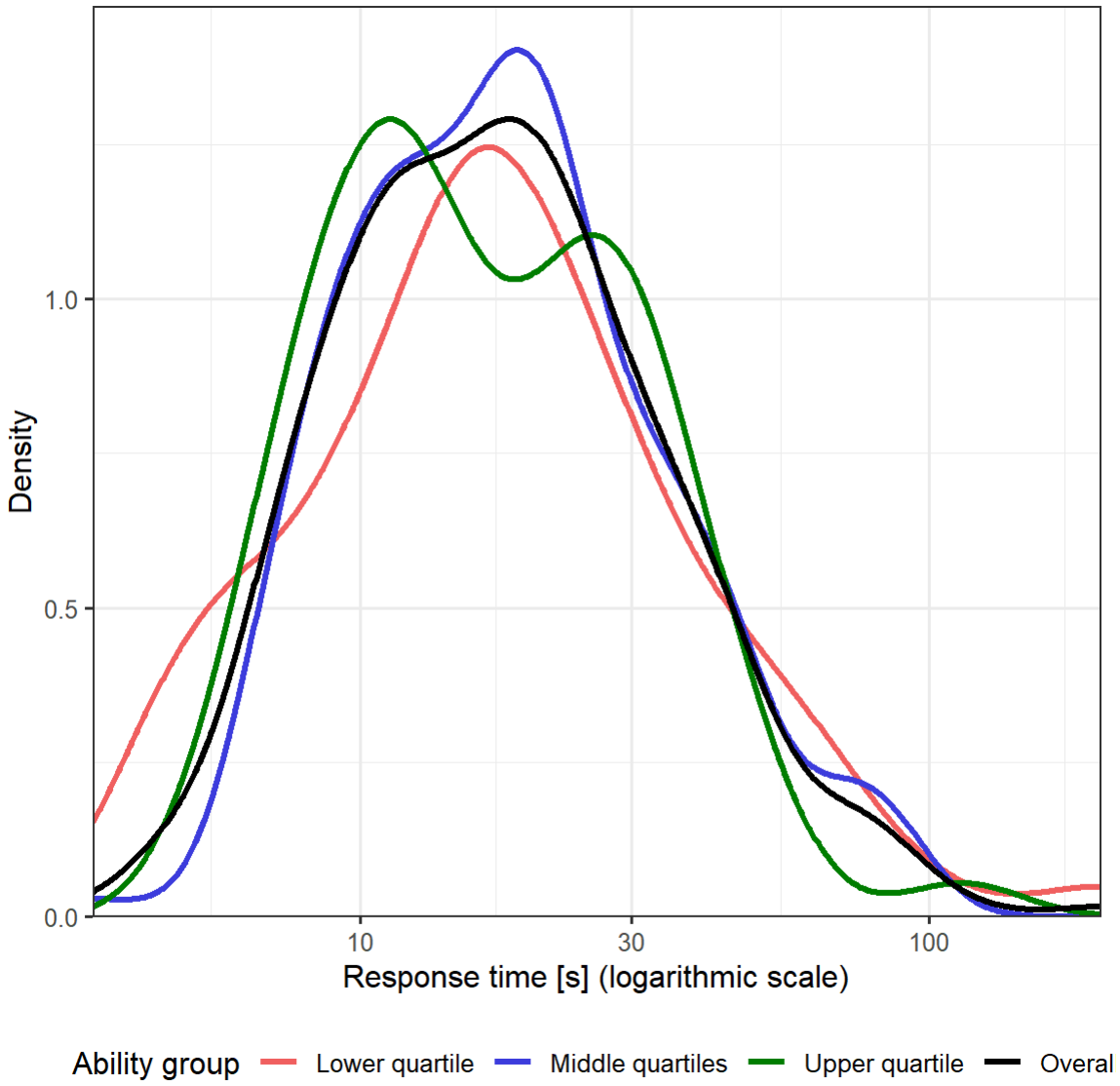
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Krokodyl	17	32	9	58	32.7%	34.4%	42.9%	34.9%
Rekin	12	20	4	36	23.1%	21.5%	19.0%	21.7%
Żółw	23	41	8	72	44.2%	44.1%	38.1%	43.4%
Overall	52	93	21	166	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



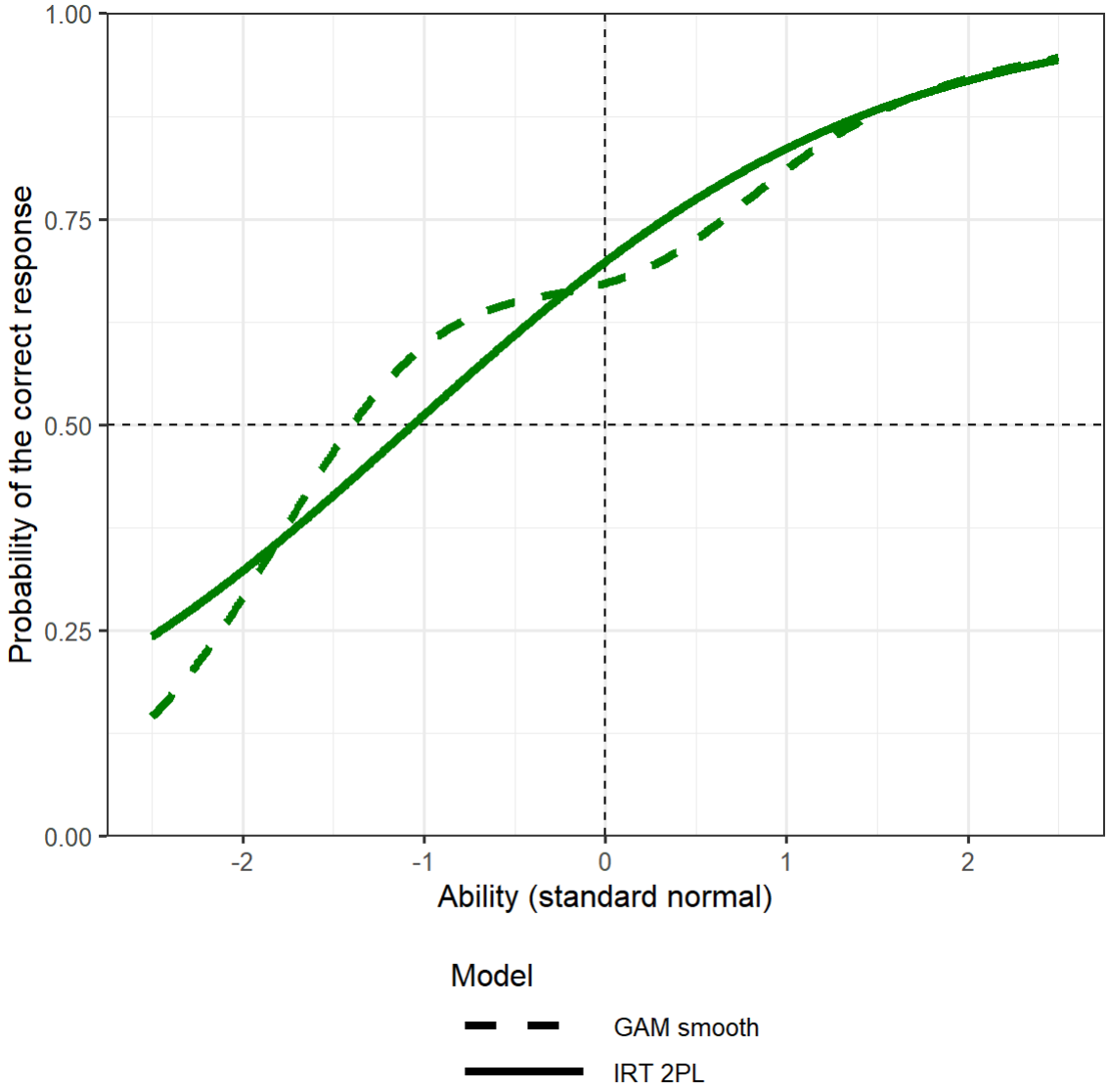
10.2.26.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.4	10.3	16.8	26.6	199.8	2.9	0.8
Middle quartiles	3.4	11.2	18.1	26.5	87.6	2.9	0.6
Upper quartile	5.6	10.4	15.3	27.0	114.6	2.8	0.6
Overall	3.4	10.8	17.1	27.0	199.8	2.9	0.7

10.2.27 S090075 (TIMSS)

10.2.27.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	67.50%	281	LLM teachers' selection	MC	0.30	0.24	0.79	−1.06	0.0(

10.2.27.2 Gender DIF:

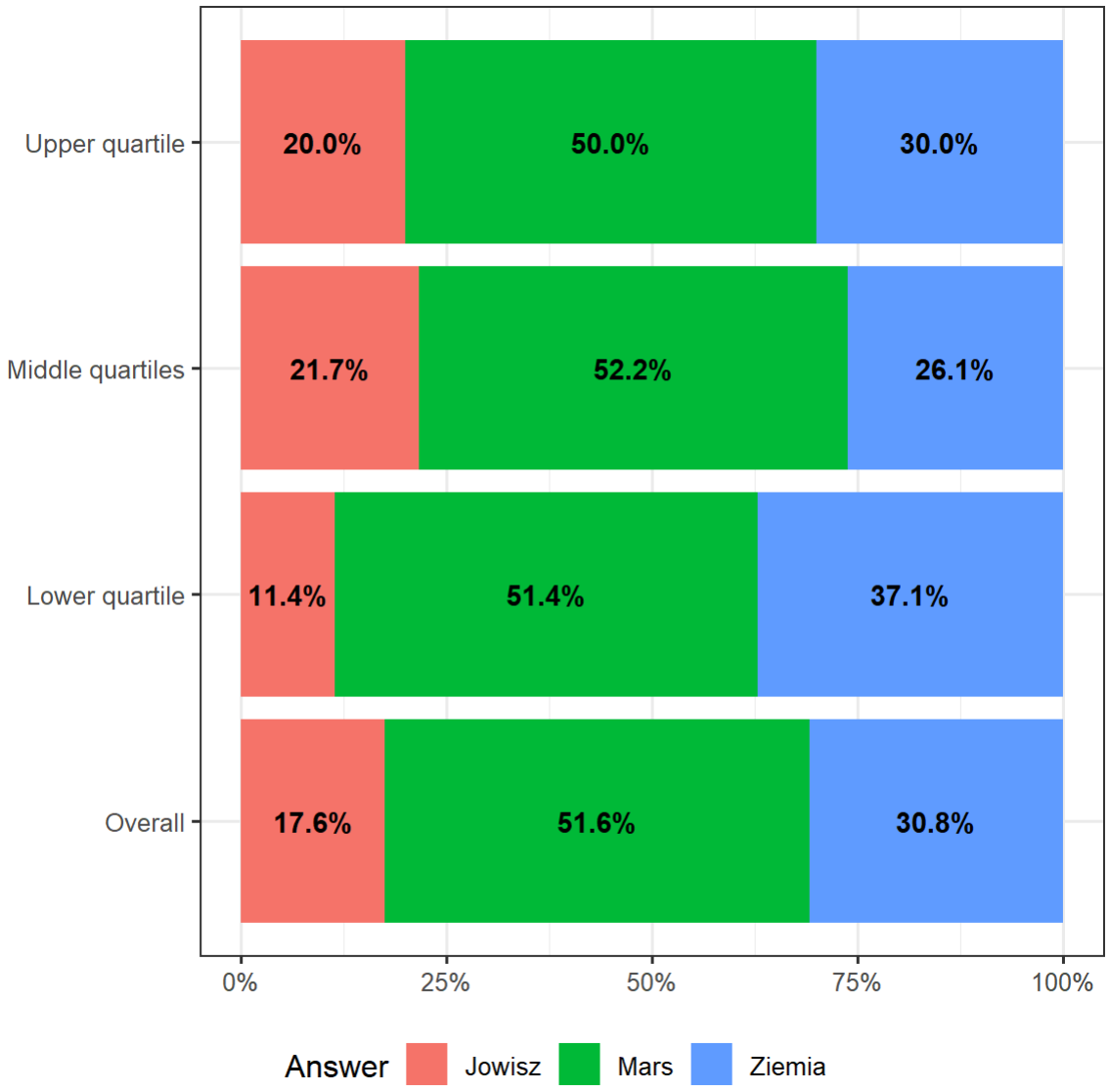
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.15	0.926	1.000	0.001	N/A

10.2.27.3 Frequency of answers:

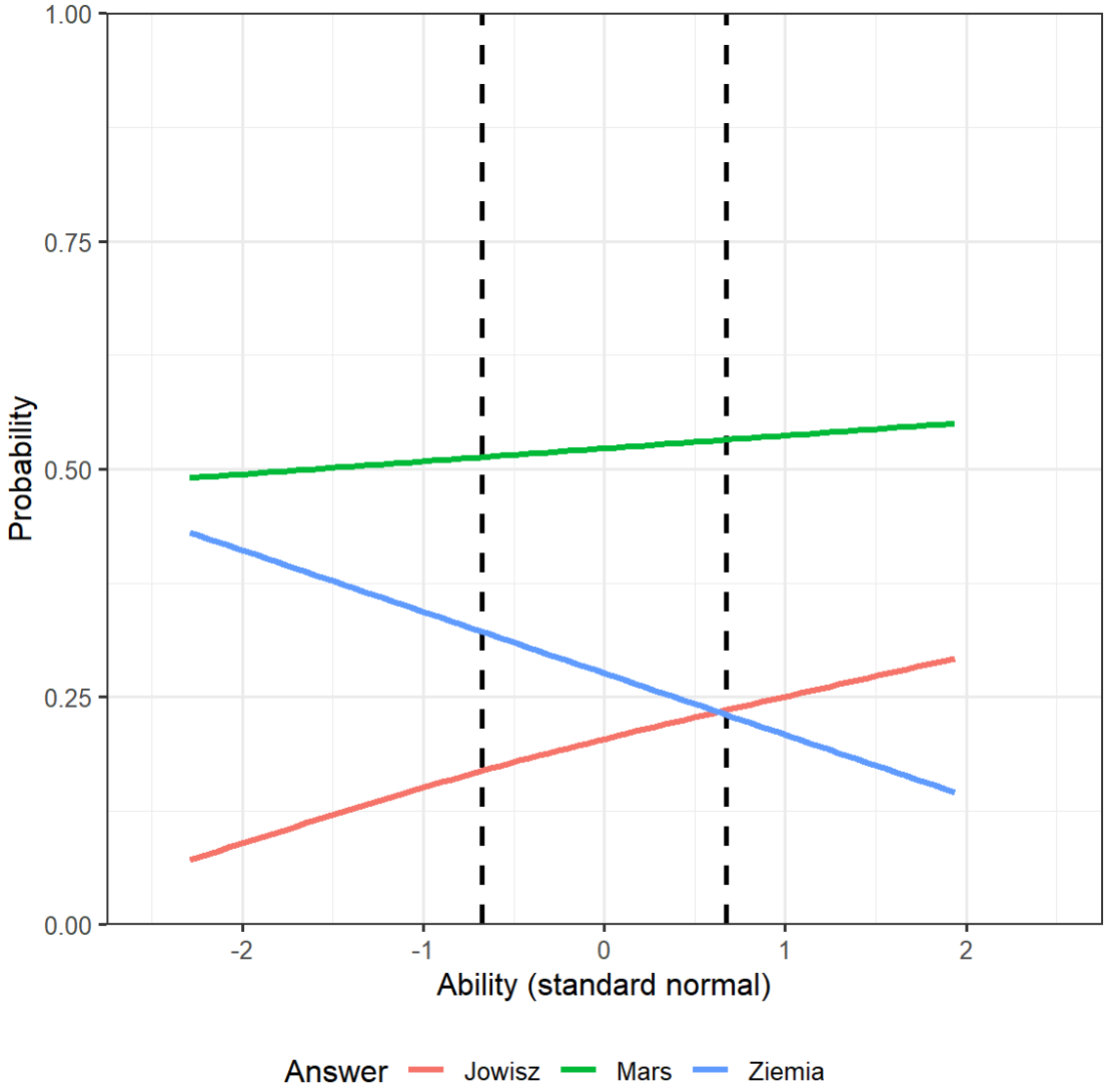
Answer	Frequency	Percent
Jowisz	16	5.7%
Mars	47	16.8%
Merkury (correct)	189	67.5%
Ziemia	28	10.0%
Overall	280	100.0%

10.2.27.4 Distractors functioning:

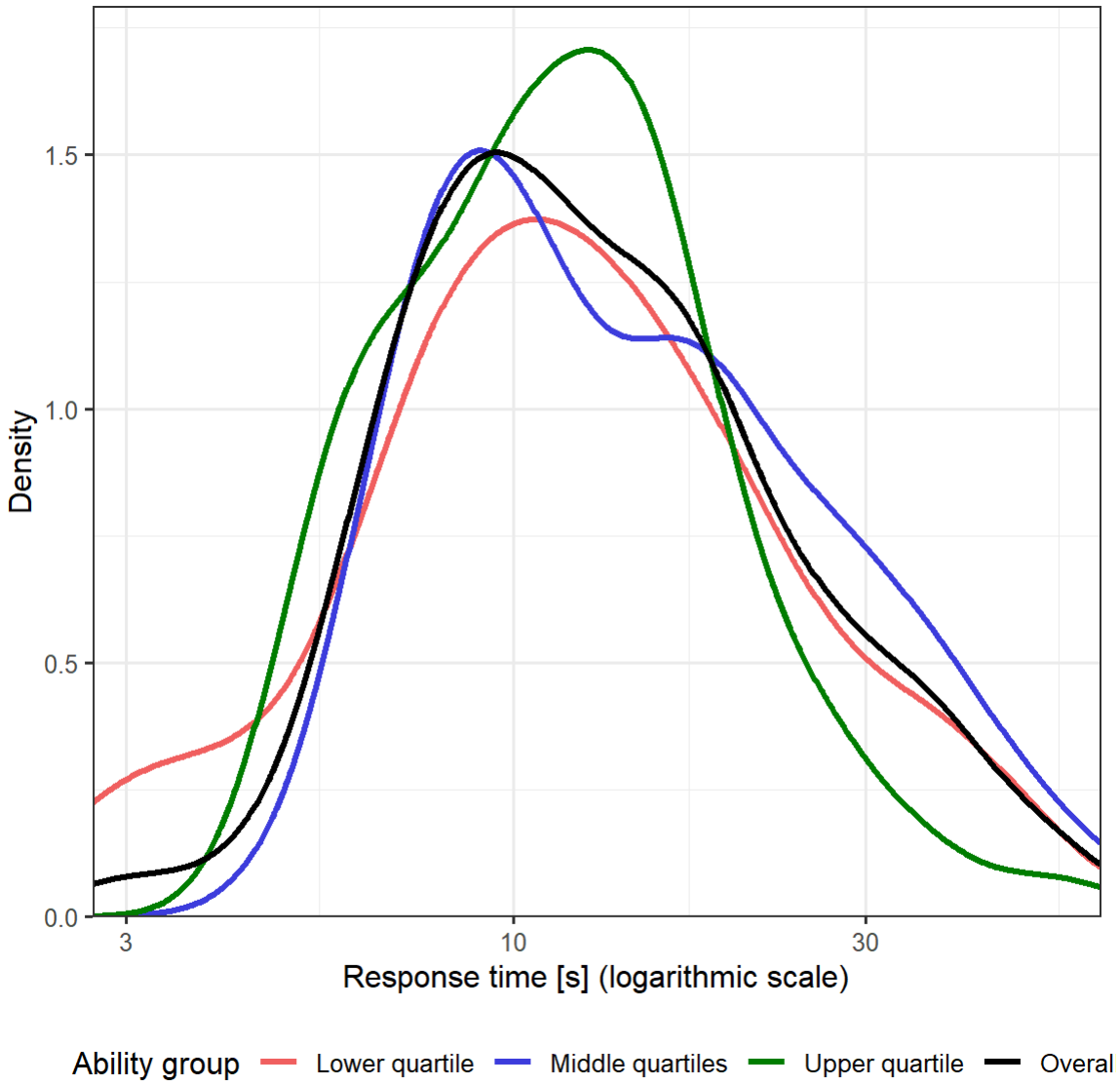
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Jowisz	4	10	2	16	11.4%	21.7%	20.0%	17.6%
Mars	18	24	5	47	51.4%	52.2%	50.0%	51.6%
Ziemia	13	12	3	28	37.1%	26.1%	30.0%	30.8%
Overall	35	46	10	91	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



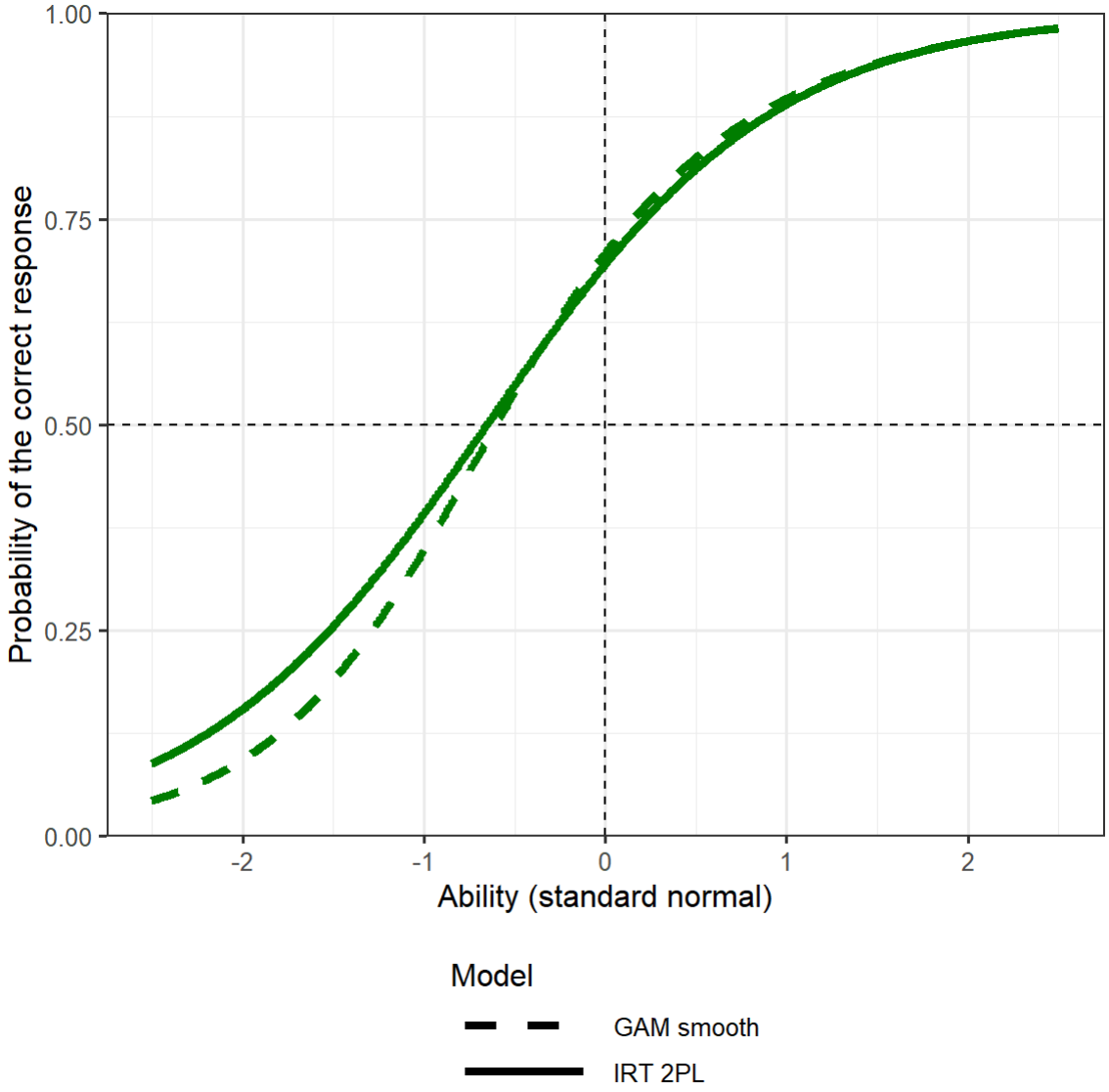
10.2.27.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	7.8	11.6	19.0	49.6	2.5	0.7
Middle quartiles	4.9	8.8	12.5	21.6	62.3	2.7	0.6
Upper quartile	5.0	7.7	11.3	15.5	55.4	2.5	0.5
Overall	2.7	8.6	12.1	19.2	62.3	2.6	0.6

10.2.28 S090020 (TIMSS)

10.2.28.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	64.80%	304	LLM teachers' selection	MC	0.48	0.34	1.26	−0.65	0.02

10.2.28.2 Gender DIF:

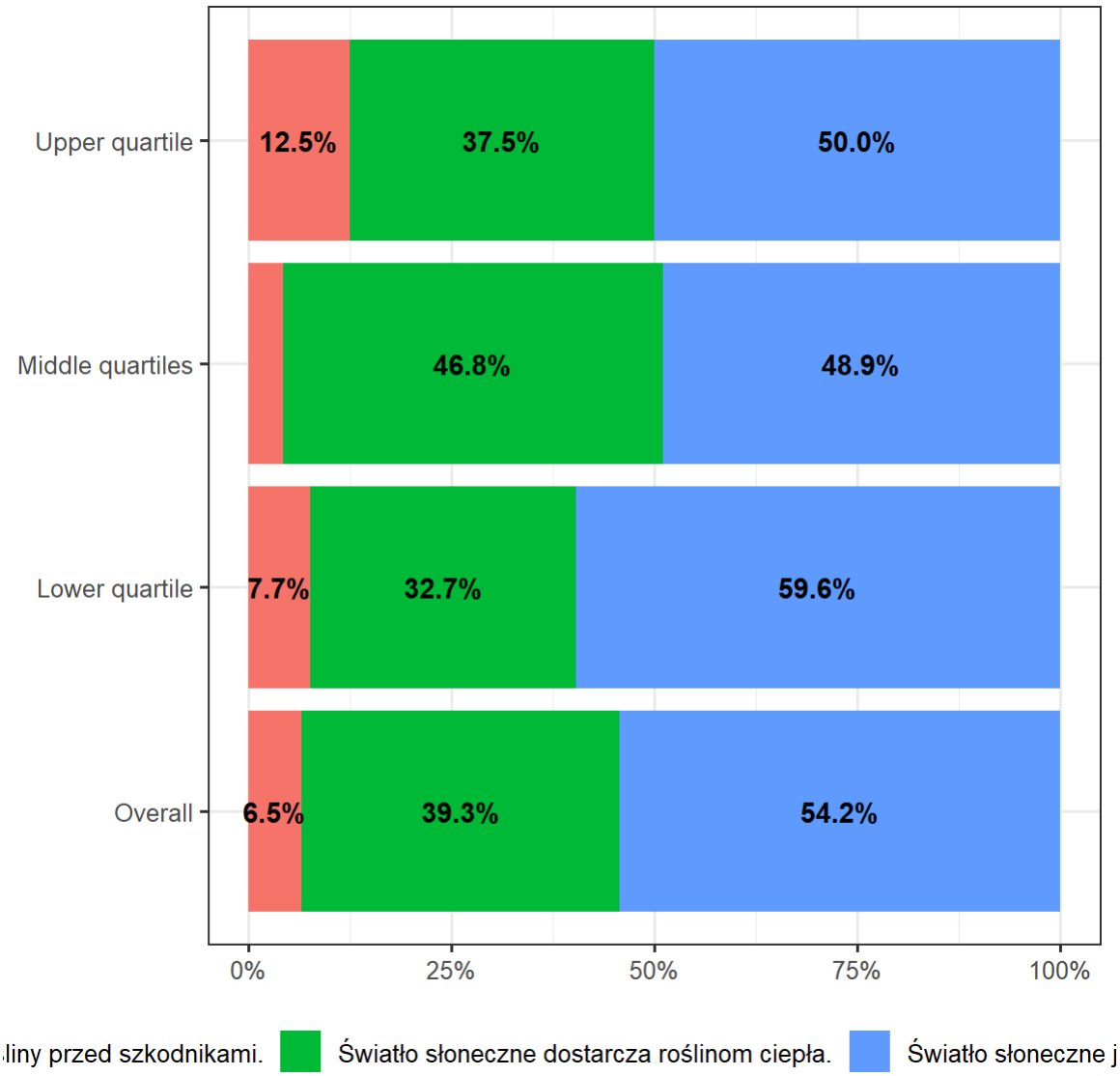
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.72	0.422	1.000	0.005	N/A

10.2.28.3 Frequency of answers:

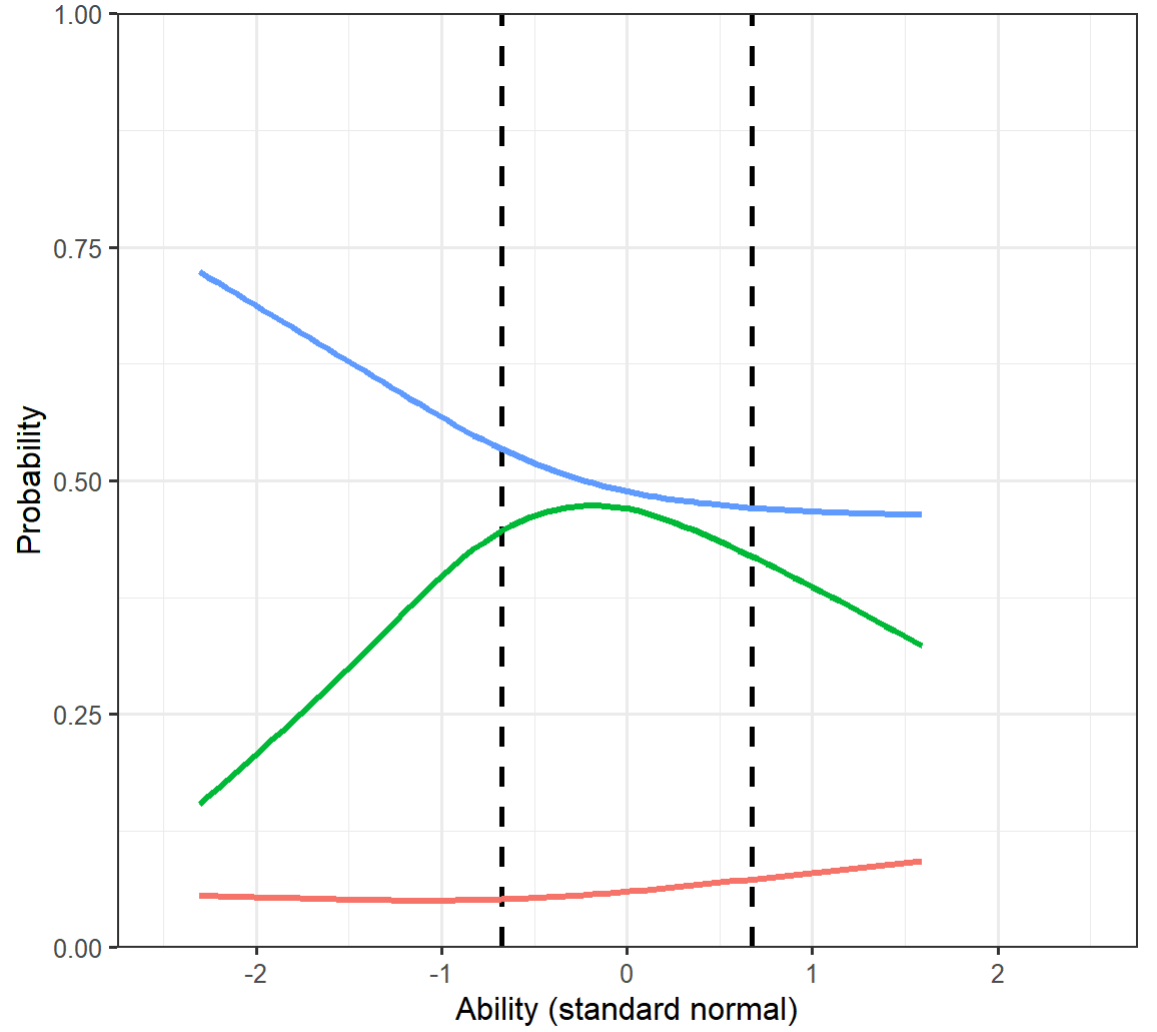
Answer	Frequency	Percent
Światło słoneczne chroni rośliny przed szkodnikami.	7	2.3%
Światło słoneczne dostarcza roślinom ciepła.	42	13.8%
Światło słoneczne jest potrzebne do procesu oddychania roślin.	58	19.1%
Światło słoneczne umożliwia roślinom fotosyntezę. (correct)	197	64.8%
Overall	304	100.0%

10.2.28.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Światło słoneczne chroni rośliny przed szkodnikami.	4	2	1	7	7.7%	4.3%	12.5%	6.5%
Światło słoneczne dostarcza roślinom ciepła.	17	22	3	42	32.7%	46.8%	37.5%	39.3%
Światło słoneczne jest potrzebne do procesu oddychania roślin.	31	23	4	58	59.6%	48.9%	50.0%	54.2%
Overall	52	47	8	107	100.0%	100.0%	100.0%	100.0%

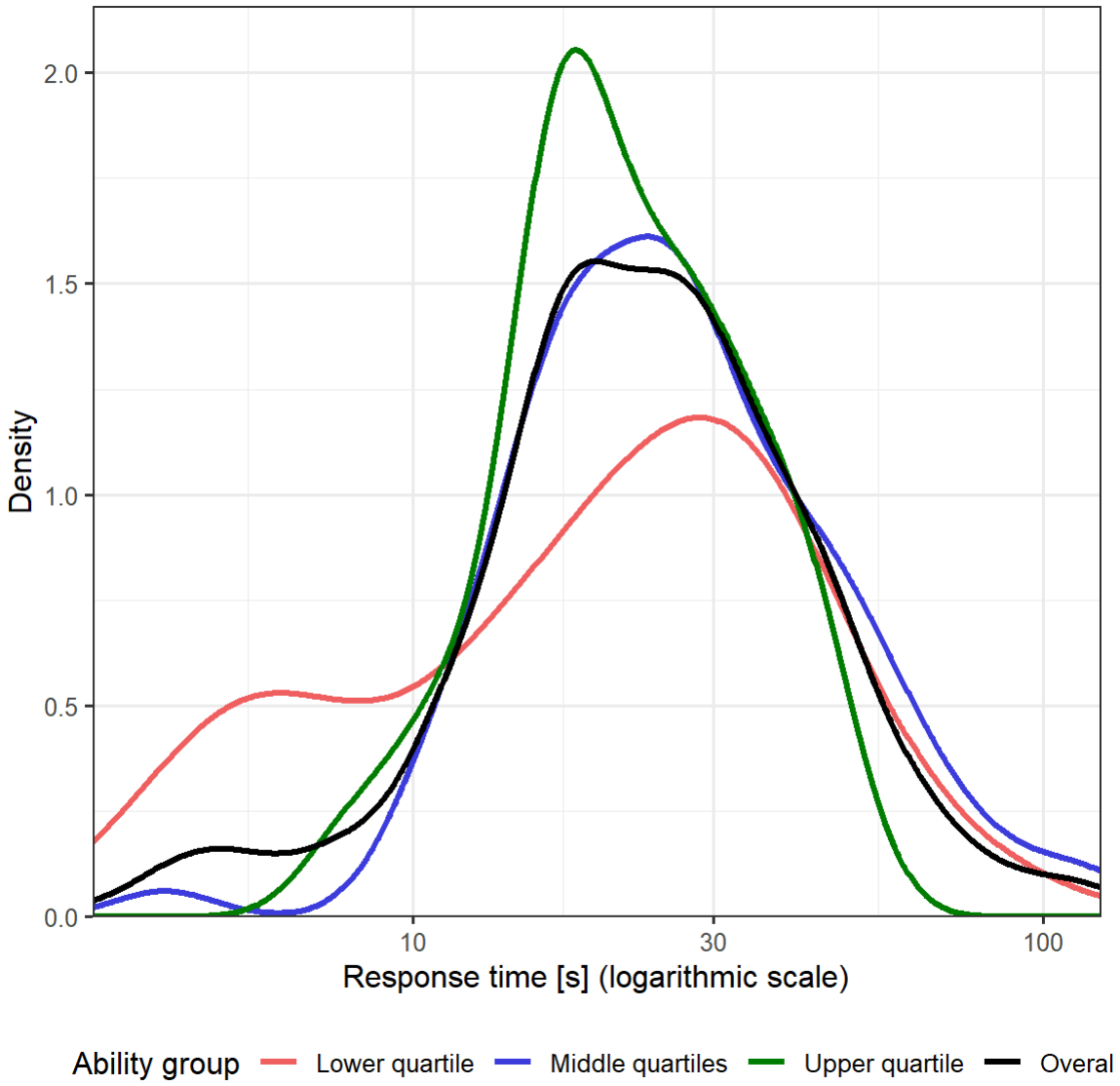



```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



liny przed szkodnikami. Światło słoneczne dostarcza roślinom ciepła. Światło słoneczne j

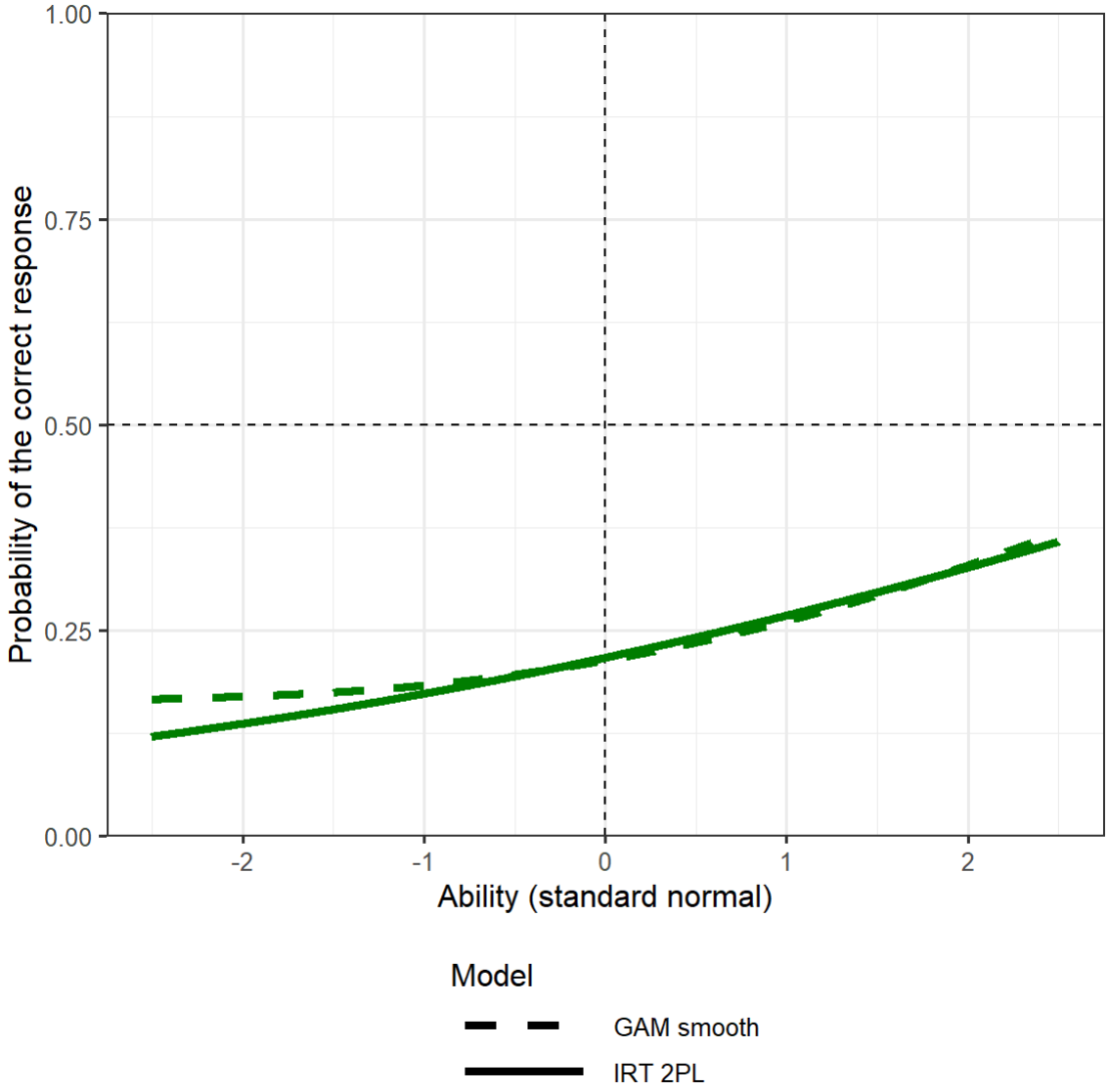
10.2.28.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	11.9	21.7	33.6	97.9	2.9	0.8
Middle quartiles	4.0	17.5	25.1	37.7	123.4	3.3	0.6
Upper quartile	7.5	16.6	20.8	29.5	49.9	3.1	0.4
Overall	3.1	16.5	23.5	34.0	123.4	3.1	0.6

10.2.29 S100009 (LLM)

10.2.29.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	22.15%	299	LLM simple setup	MC	0.09	0.04	0.28	4.59	0.015

10.2.29.2 Gender DIF:

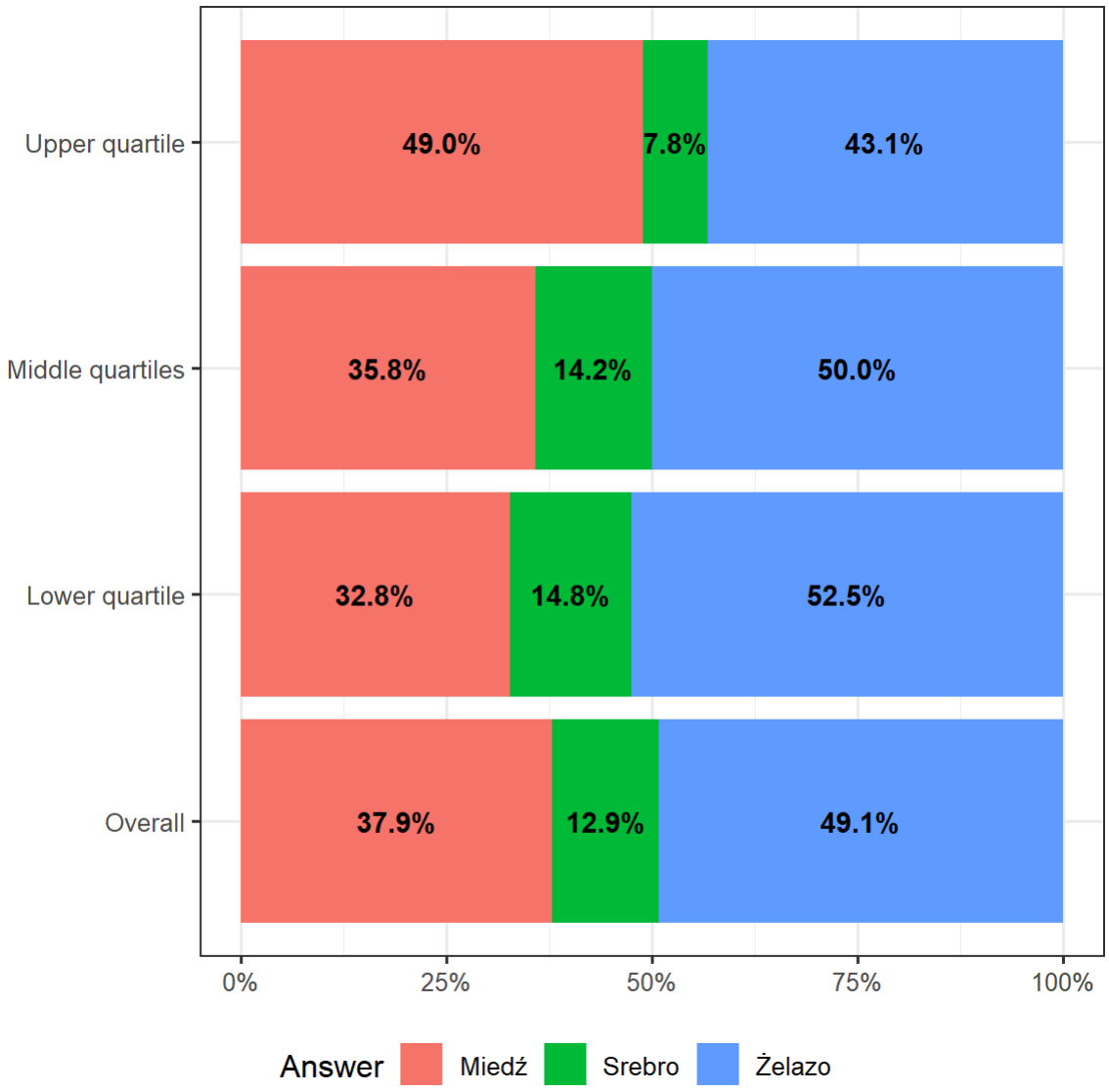
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.63	0.730	1.000	0.002	N/A

10.2.29.3 Frequency of answers:

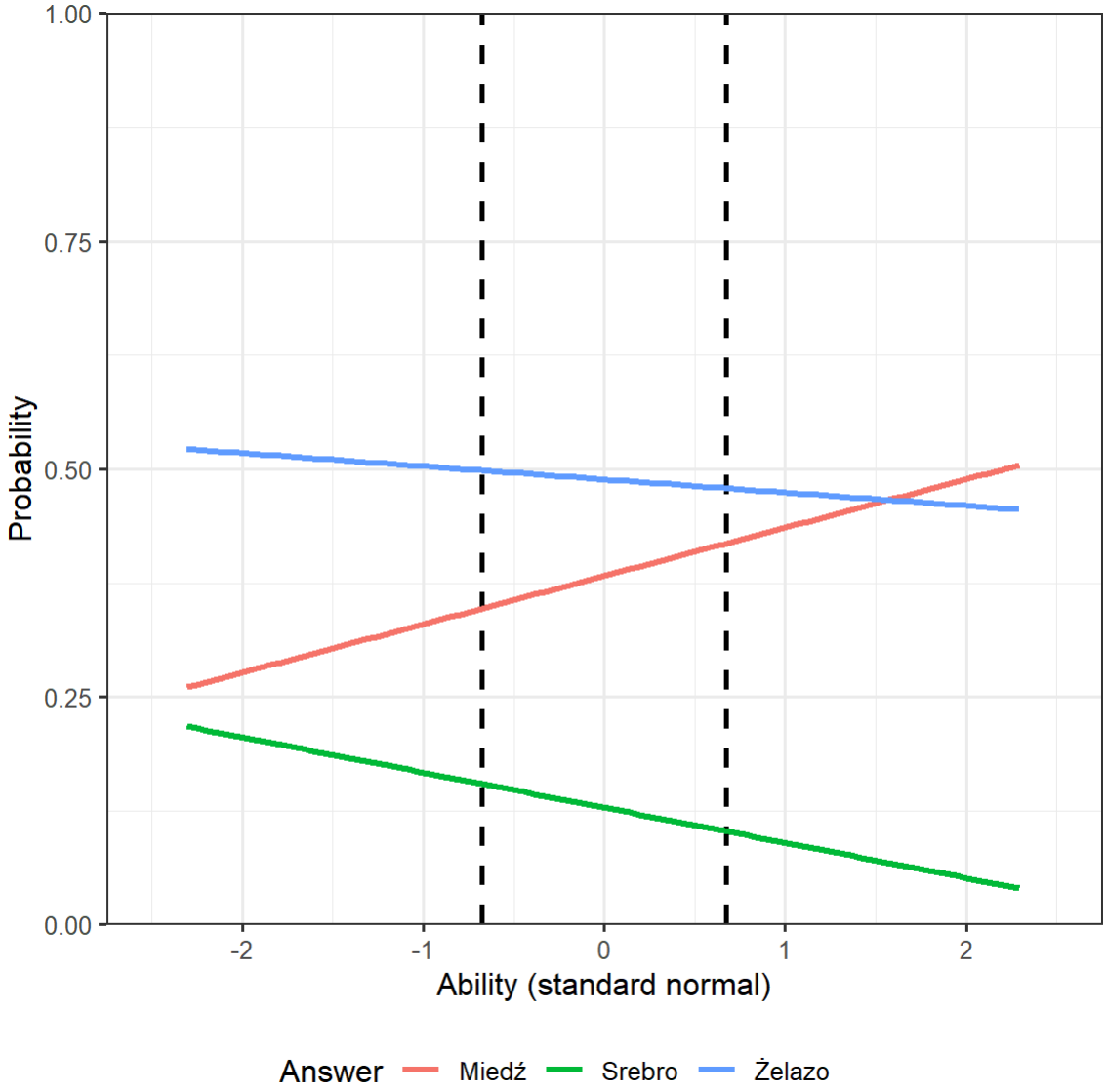
Answer	Frequency	Percent
Guma (correct)	66	22.1%
Miedź	88	29.5%
Srebro	30	10.1%
Żelazo	114	38.3%
Overall	298	100.0%

10.2.29.4 Distractors functioning:

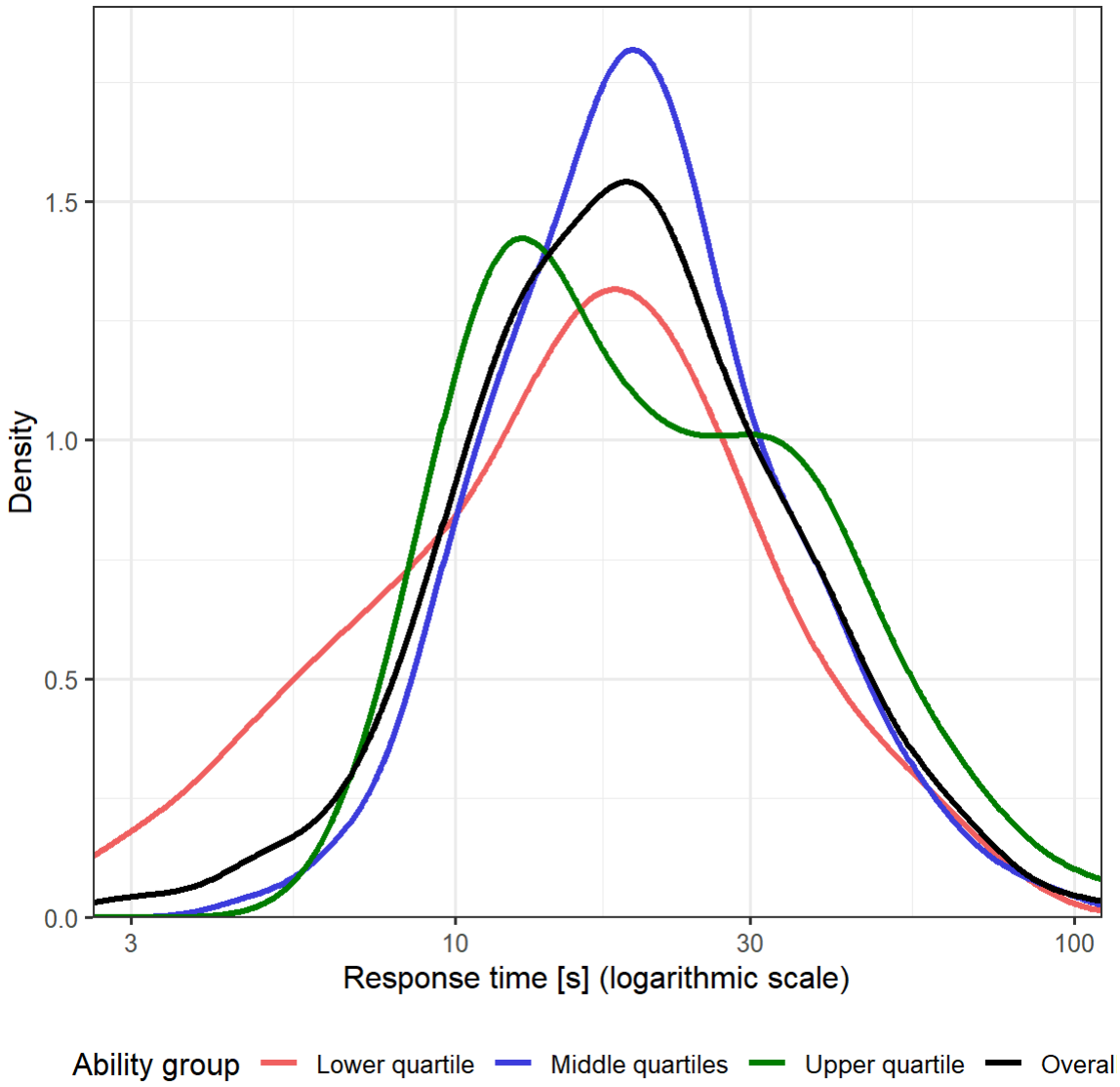
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Miedź	20	43	25	88	32.8%	35.8%	49.0%	37.9%
Srebro	9	17	4	30	14.8%	14.2%	7.8%	12.9%
Żelazo	32	60	22	114	52.5%	50.0%	43.1%	49.1%
Overall	61	120	51	232	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



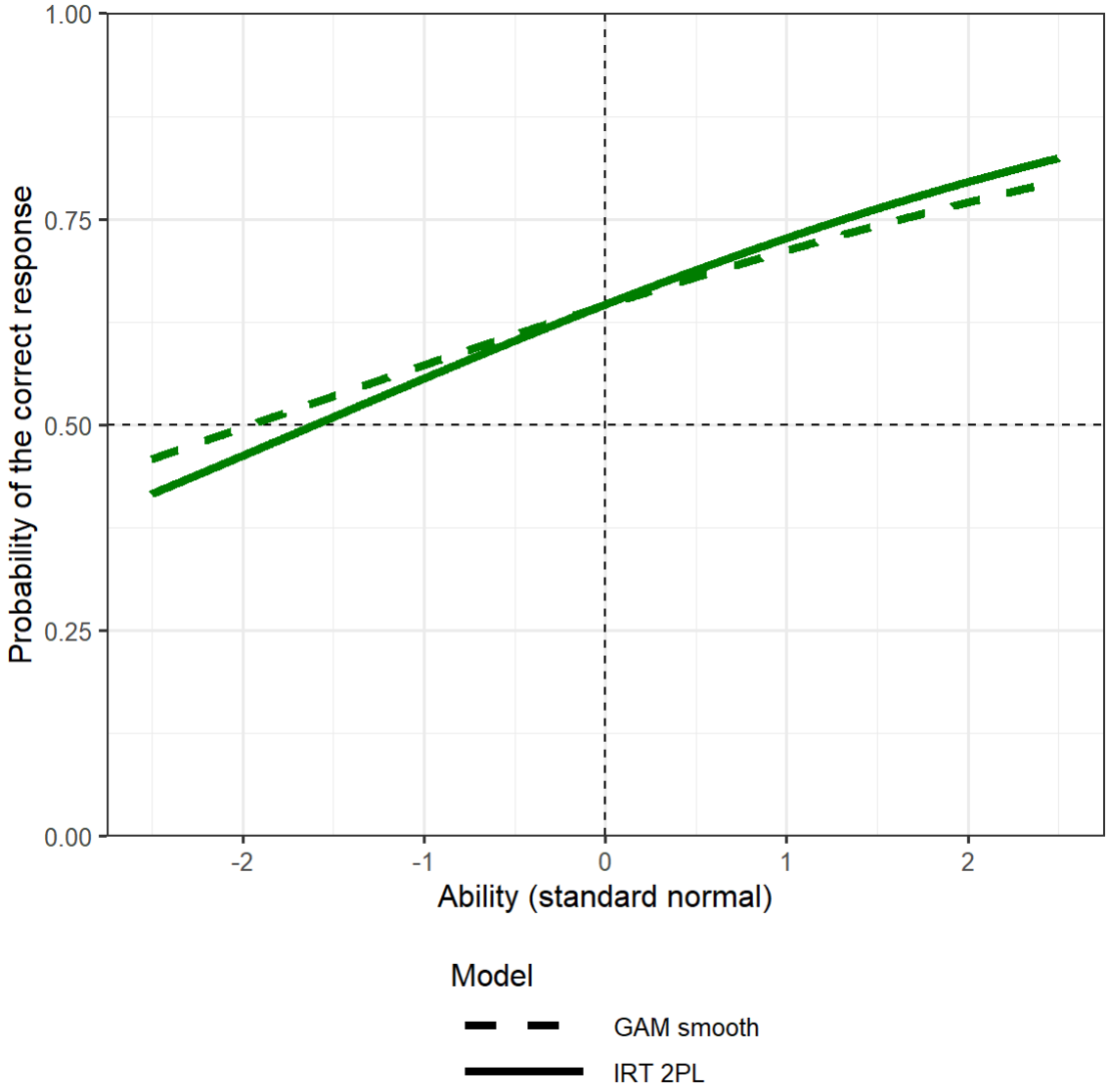
10.2.29.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.6	9.5	15.8	23.2	68.2	2.7	0.7
Middle quartiles	4.8	13.6	19.6	26.6	95.0	3.0	0.5
Upper quartile	7.4	12.0	18.7	33.9	110.5	3.0	0.6
Overall	2.6	12.7	18.4	27.7	110.5	2.9	0.6

10.2.30 S051111 (TIMSS)

10.2.30.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	64.63%	294	TIMSS	MC	0.14	0.20	0.38	−1.60	0.000

10.2.30.2 Gender DIF:

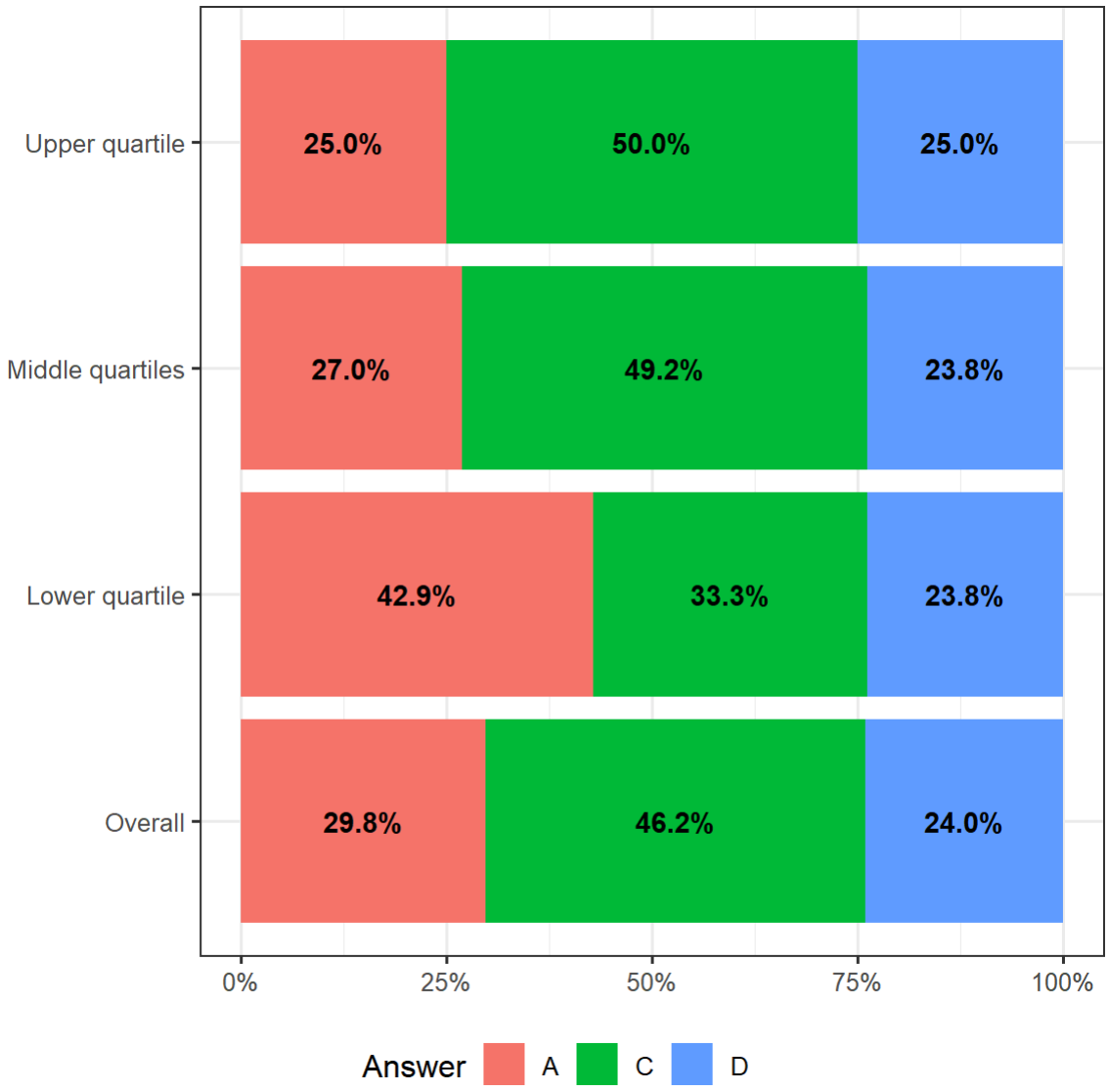
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.22	0.544	1.000	0.004	N/A

10.2.30.3 Frequency of answers:

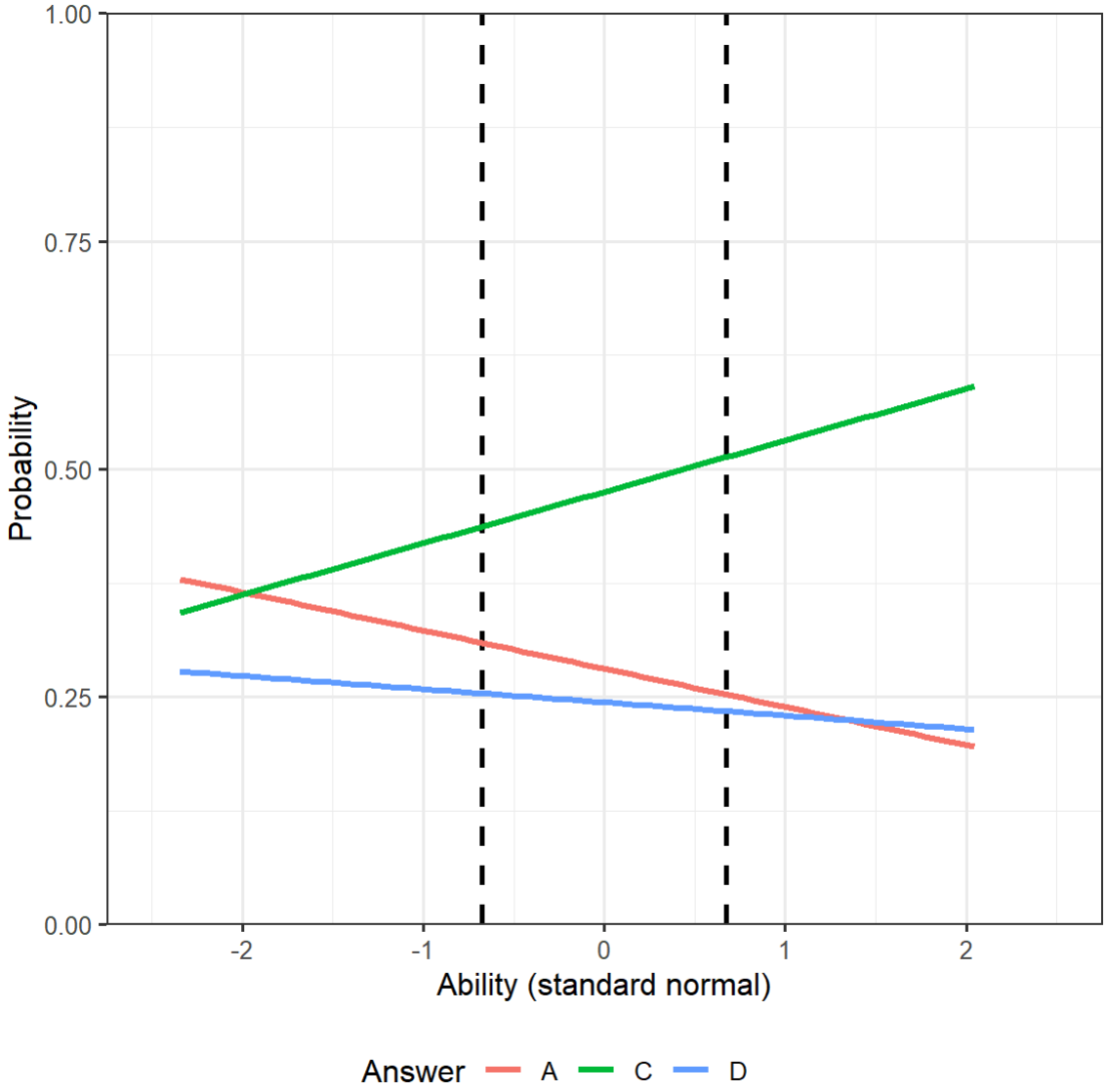
Answer	Frequency	Percent
A	31	10.5%
B (correct)	190	64.6%
C	48	16.3%
D	25	8.5%
Overall	294	100.0%

10.2.30.4 Distractors functioning:

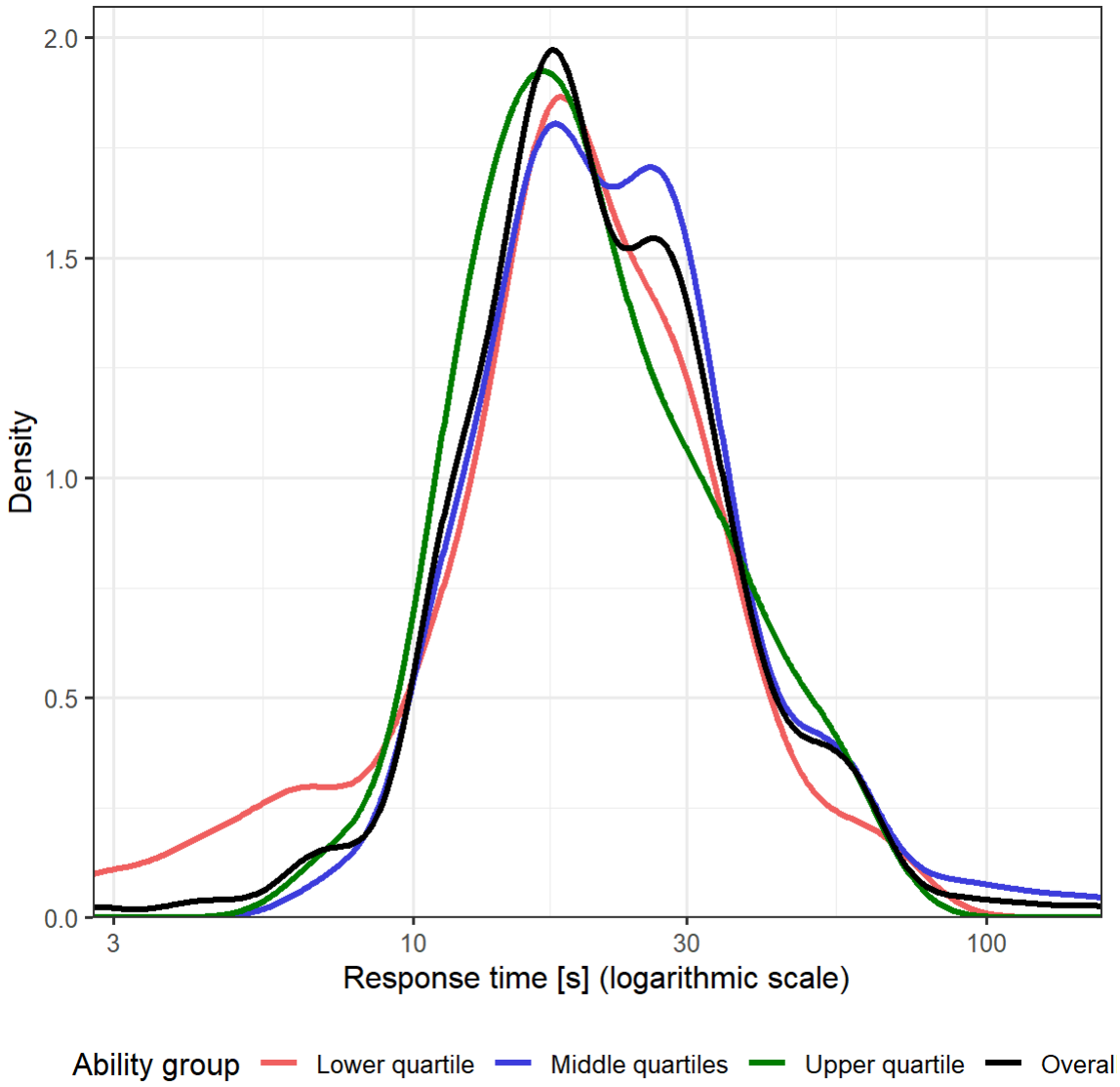
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
A	9	17	5	31	42.9%	27.0%	25.0%	29.8%
C	7	31	10	48	33.3%	49.2%	50.0%	46.2%
D	5	15	5	25	23.8%	23.8%	25.0%	24.0%
Overall	21	63	20	104	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



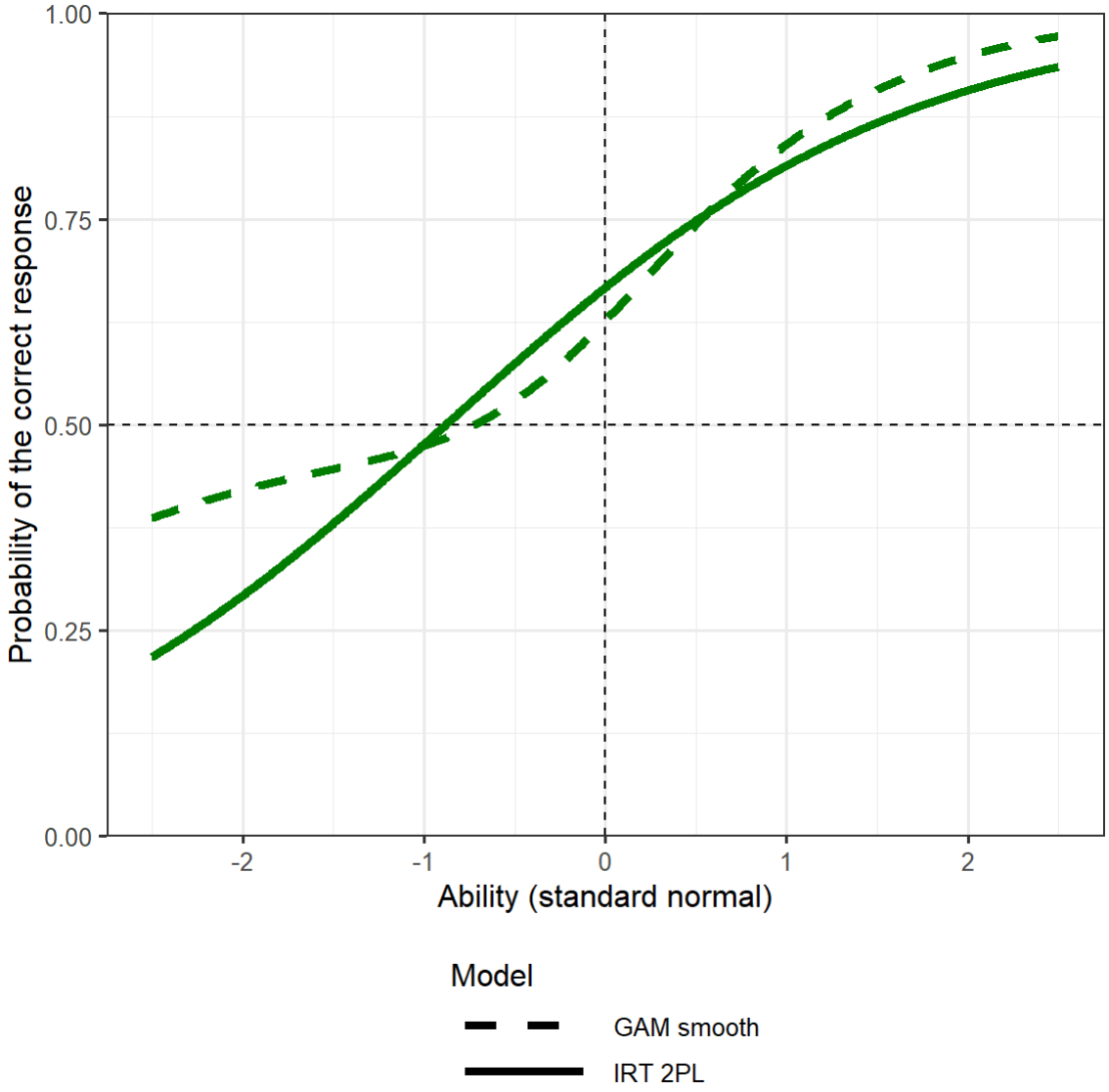
10.2.30.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.8	14.1	17.9	26.7	64.0	2.9	0.6
Middle quartiles	6.7	16.0	21.1	30.0	158.6	3.1	0.5
Upper quartile	7.1	14.4	19.2	27.9	59.8	3.0	0.5
Overall	2.8	15.4	19.9	28.5	158.6	3.0	0.5

10.2.31 S090084 (TIMSS)

10.2.31.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	64.47%	304	LLM teachers' selection	MC	0.31	0.17	0.79	−0.88	0.0

10.2.31.2 Gender DIF:

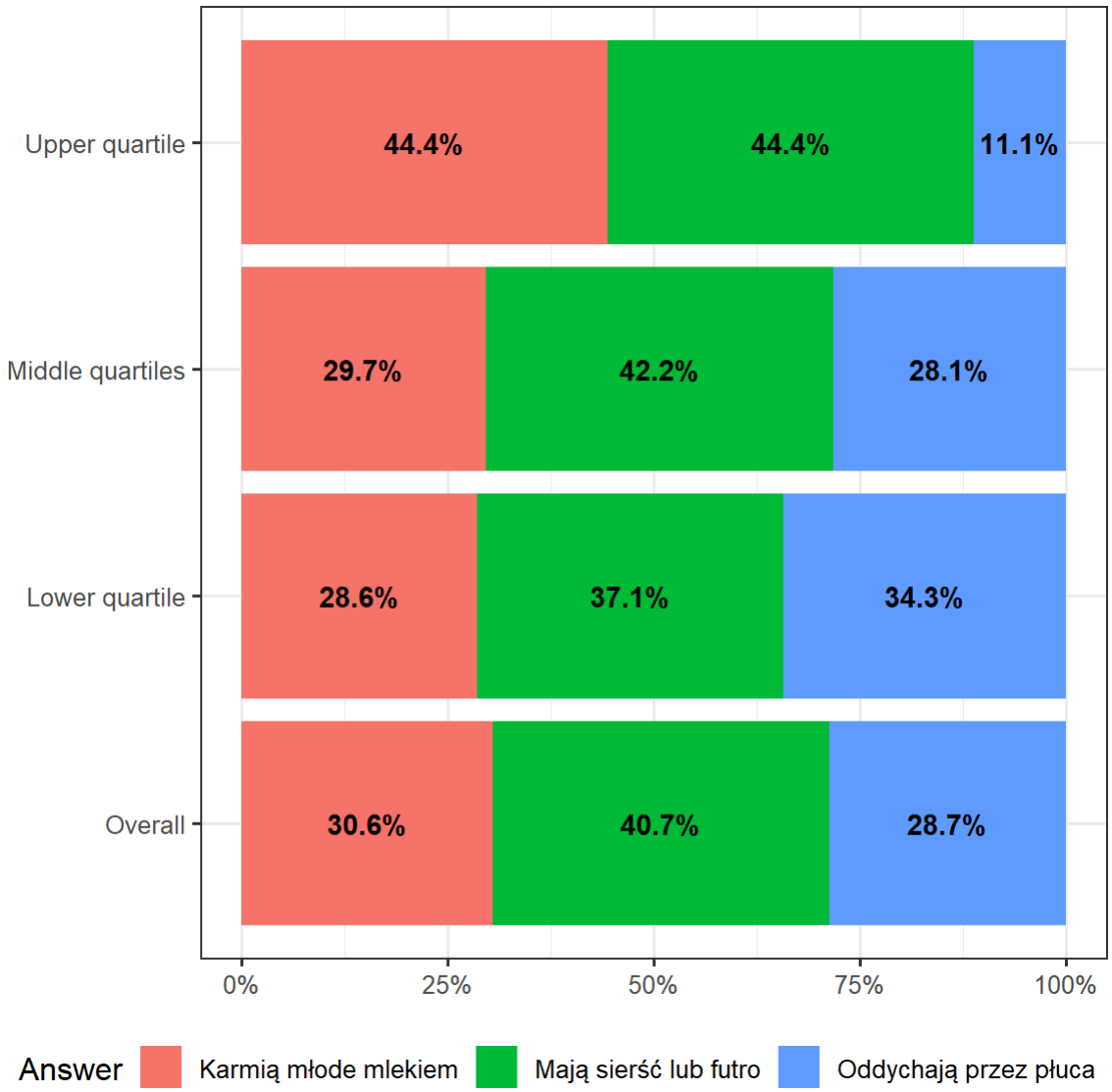
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.34	0.844	1.000	0.001	N/A

10.2.31.3 Frequency of answers:

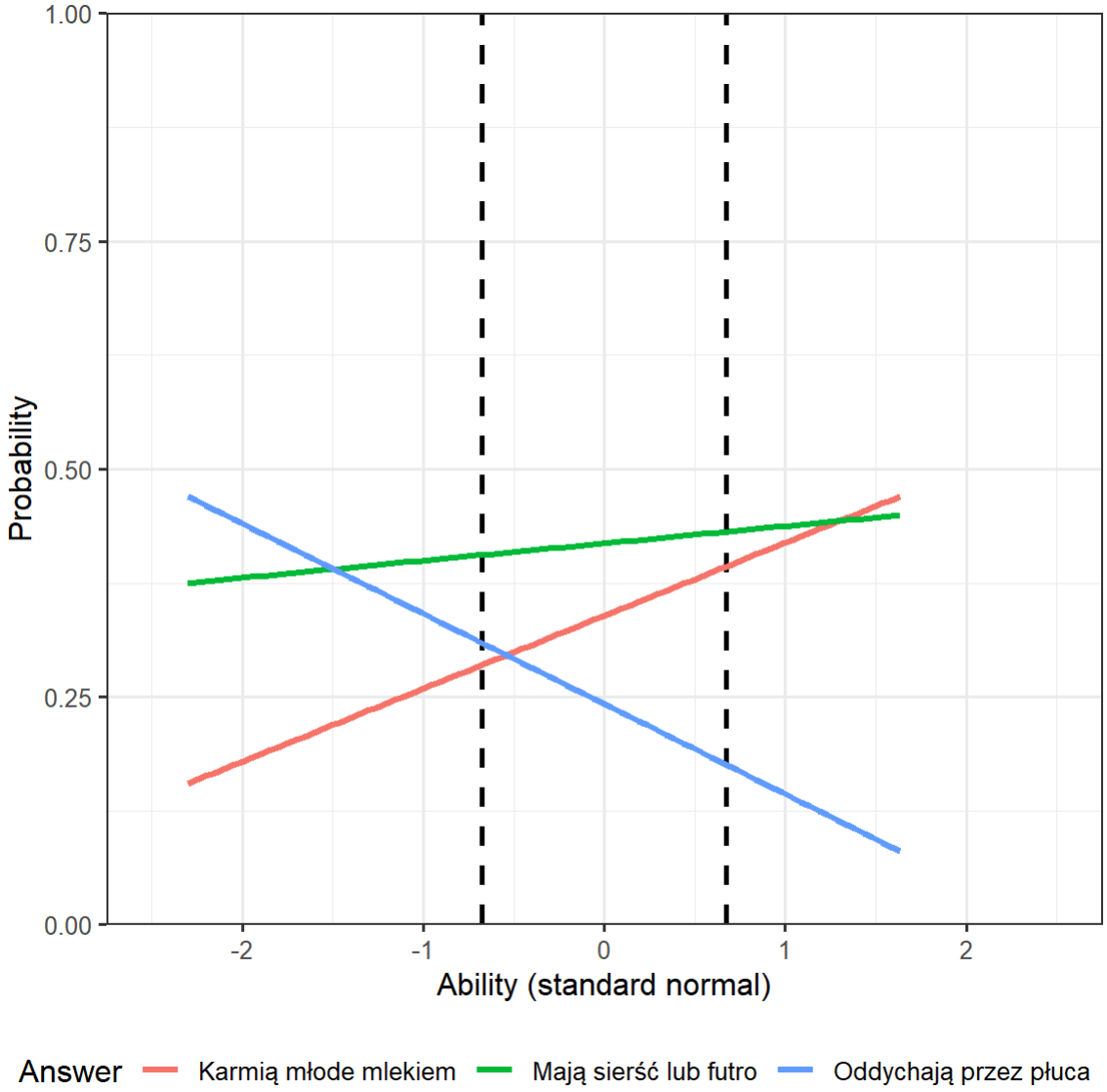
Answer	Frequency	Percent
Karmią młode mlekiem	33	10.9%
Mają sierść lub futro	44	14.5%
Oddychają przez płuca	31	10.2%
Składają jajka (correct)	196	64.5%
Overall	304	100.0%

10.2.31.4 Distractors functioning:

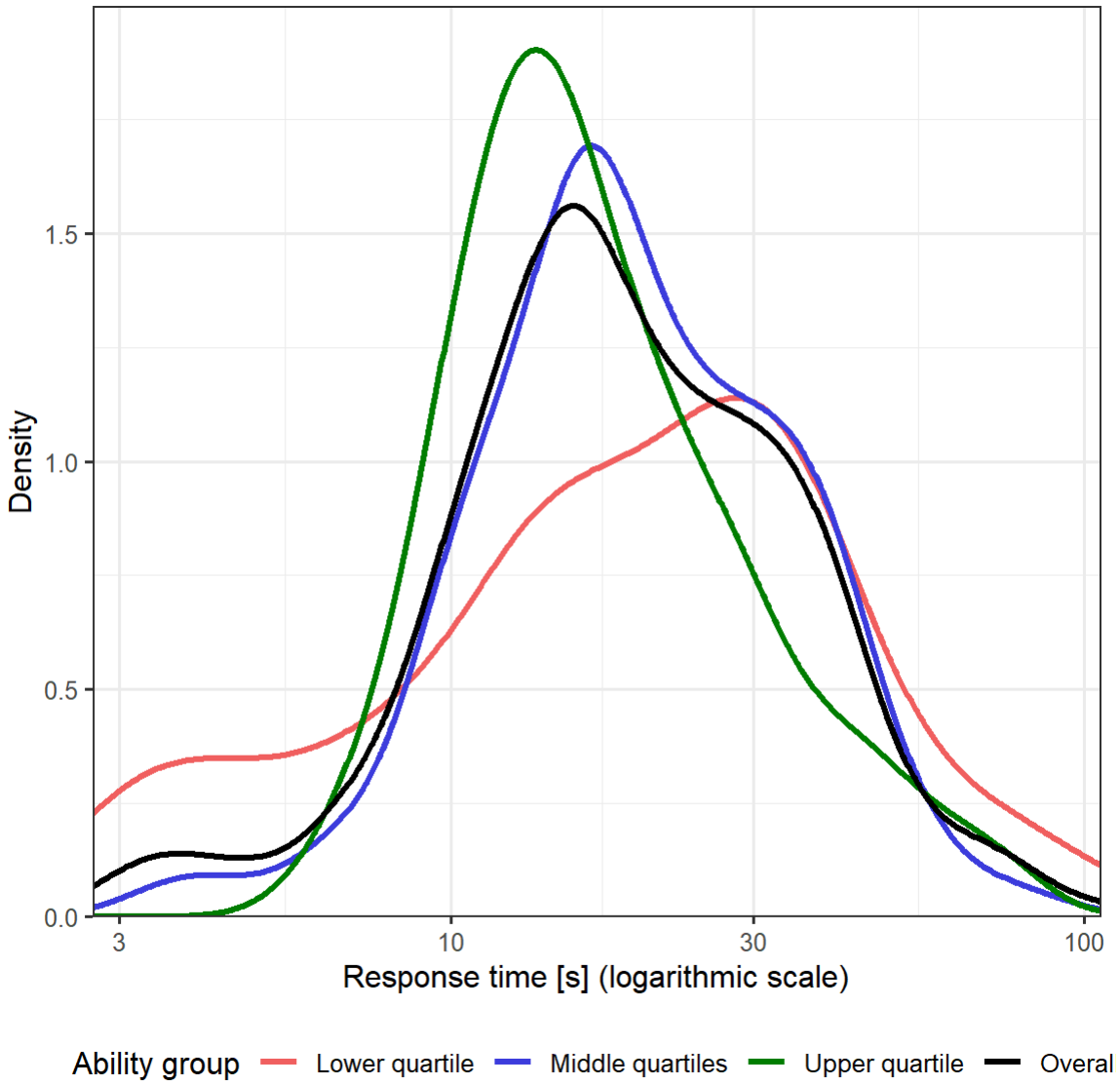
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Karmią młode mlekiem	10	19	4	33	28.6%	29.7%	44.4%	30.6%
Mają sierść lub futro	13	27	4	44	37.1%	42.2%	44.4%	40.7%
Oddychają przez płuca	12	18	1	31	34.3%	28.1%	11.1%	28.7%
Overall	35	64	9	108	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



10.2.31.5 Answering time:



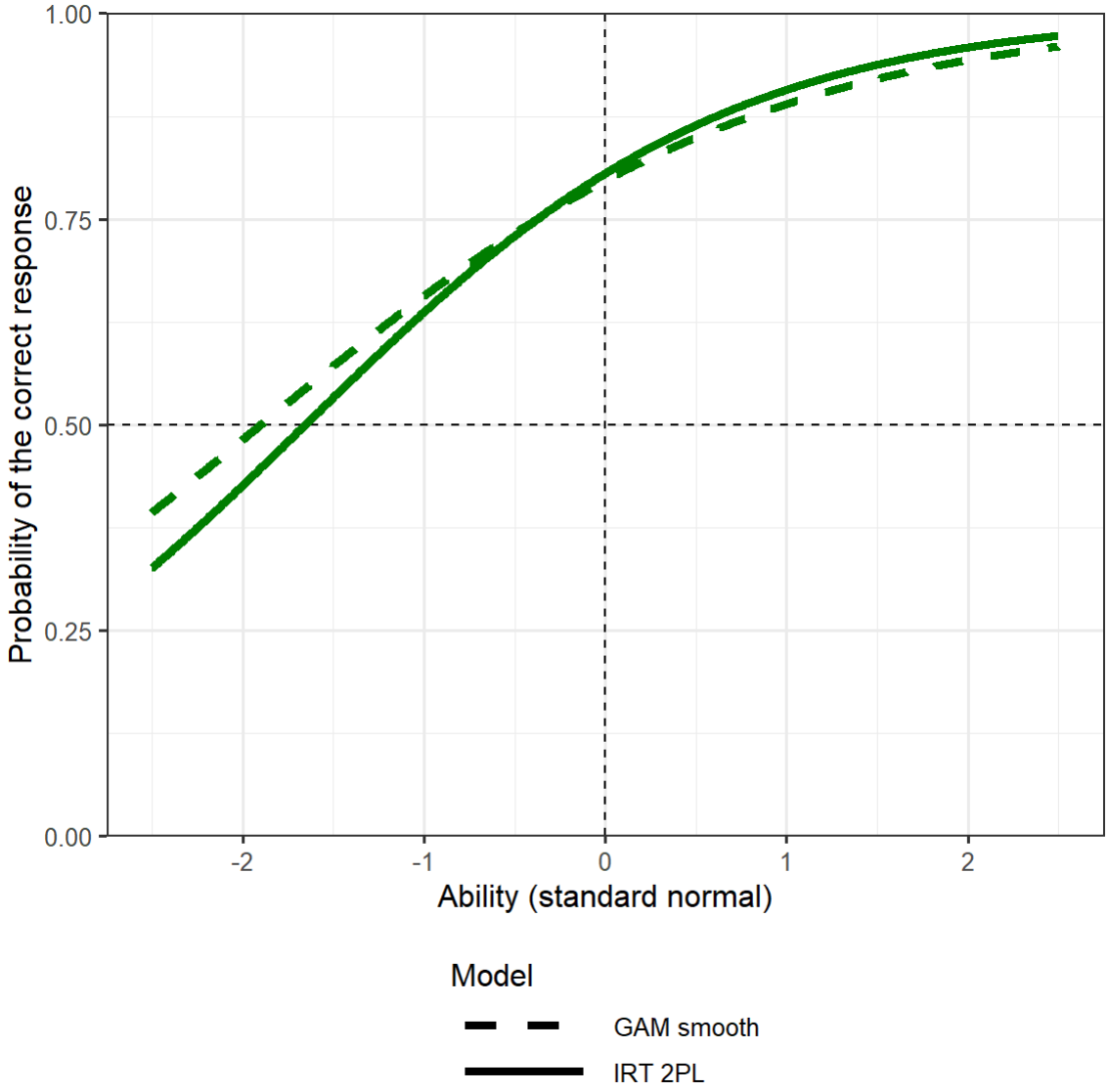
	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	11.9	20.1	32.9	106.0	2.9	0.8
Middle quartiles	3.4	13.7	18.1	29.1	84.6	3.0	0.6
Upper quartile	6.6	11.8	15.2	23.1	72.1	2.8	0.5
Overall	2.7	12.7	17.8	28.8	106.0	2.9	0.6

10.2.32 SP71128 (LLM)

10.2.32.1 Item fit:

```
## Warning in newton(lsp = lsp, X = G$X, y = G$y, Eb = G$Eb, UrS = G$UrS, L = G$L,
## : Fitting terminated with step failure - check results carefully
```

```
## Warning in newton(lsp = lsp, X = G$X, y = G$y, Eb = G$Eb, UrS = G$UrS, L = G$L,
## : Fitting terminated with step failure - check results carefully
```



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	78.23%	294	TIMSS	MC	0.28	0.32	0.86	−1.65	0.013

10.2.32.2 Gender DIF:

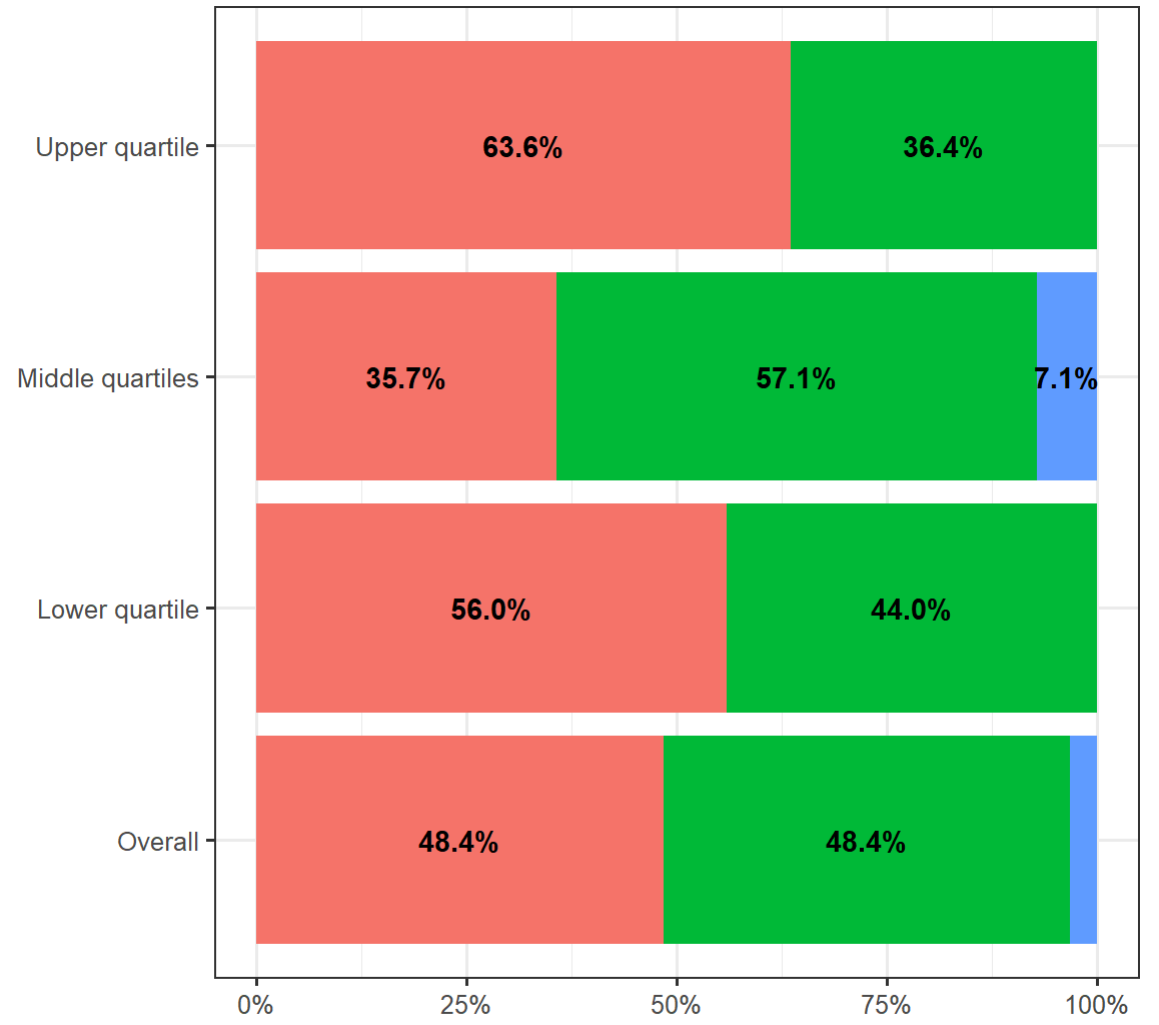
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.68	0.710	1.000	0.002	N/A

10.2.32.3 Frequency of answers:

Answer	Frequency	Percent
Energia elektryczna zamienia się w energię świetlną. (correct)	230	78.2%
Energia kinetyczna zamienia się w energię świetlną.	31	10.5%
Energia świetlna zamienia się w energię elektryczną.	31	10.5%
Energia świetlna zamienia się w energię kinetyczną.	2	0.7%
Overall	294	100.0%

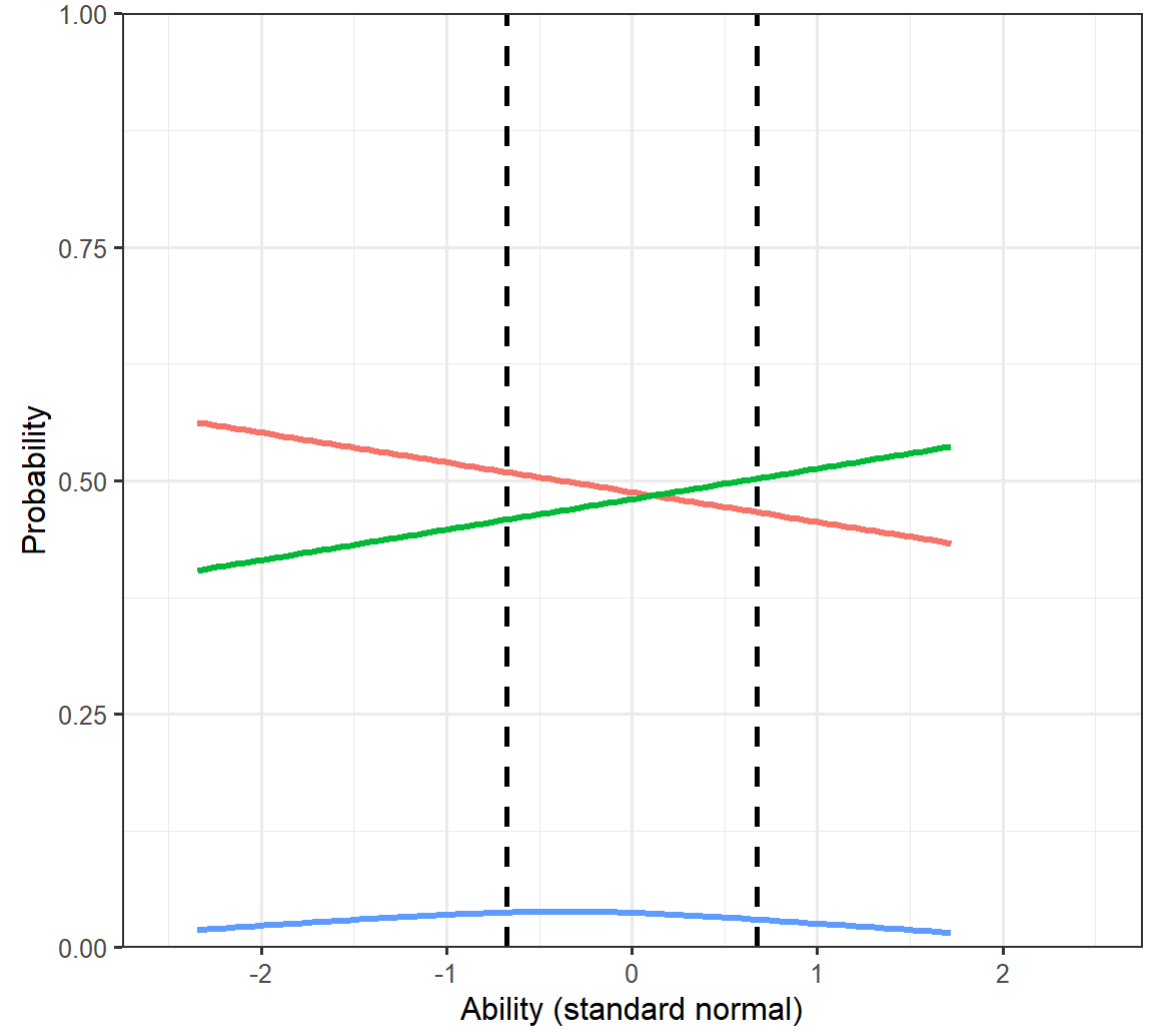
10.2.32.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Energia kinetyczna zamienia się w energię świetlną.	14	10	7	31	56.0%	35.7%	63.6%	48.4%
Energia świetlna zamienia się w energię elektryczną.	11	16	4	31	44.0%	57.1%	36.4%	48.4%
Overall	25	28	11	64	100.0%	100.0%	100.0%	100.0%
Energia świetlna zamienia się w energię kinetyczną.	0	2	0	2	0.0%	7.1%	0.0%	3.1%



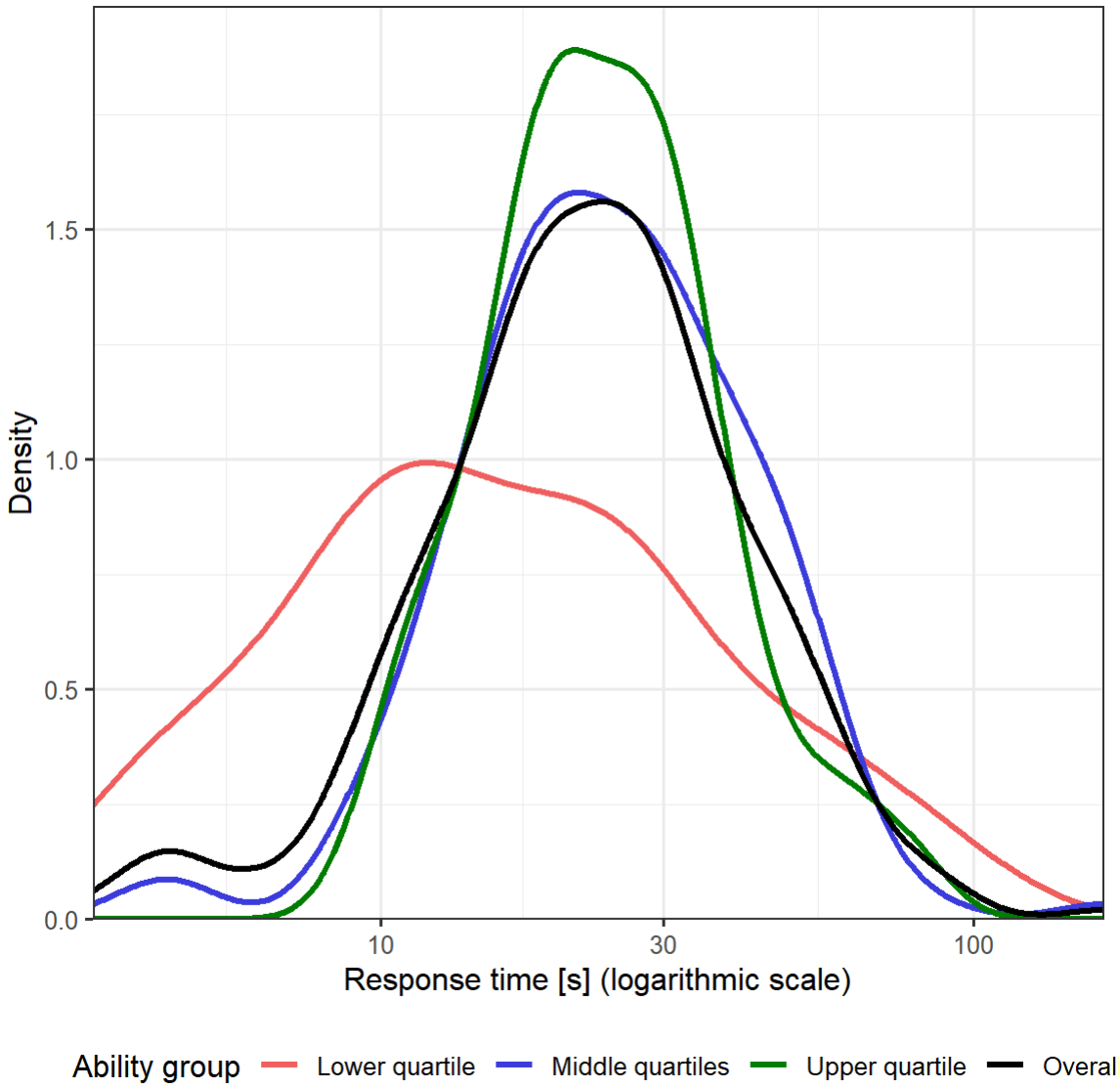
ıia się w energię świetlną. Energia świetlna zamienia się w energię elektryczną. Energia ś

```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



ıia się w energię świetlną. Energia świetlna zamienia się w energię elektryczną. Energia ś

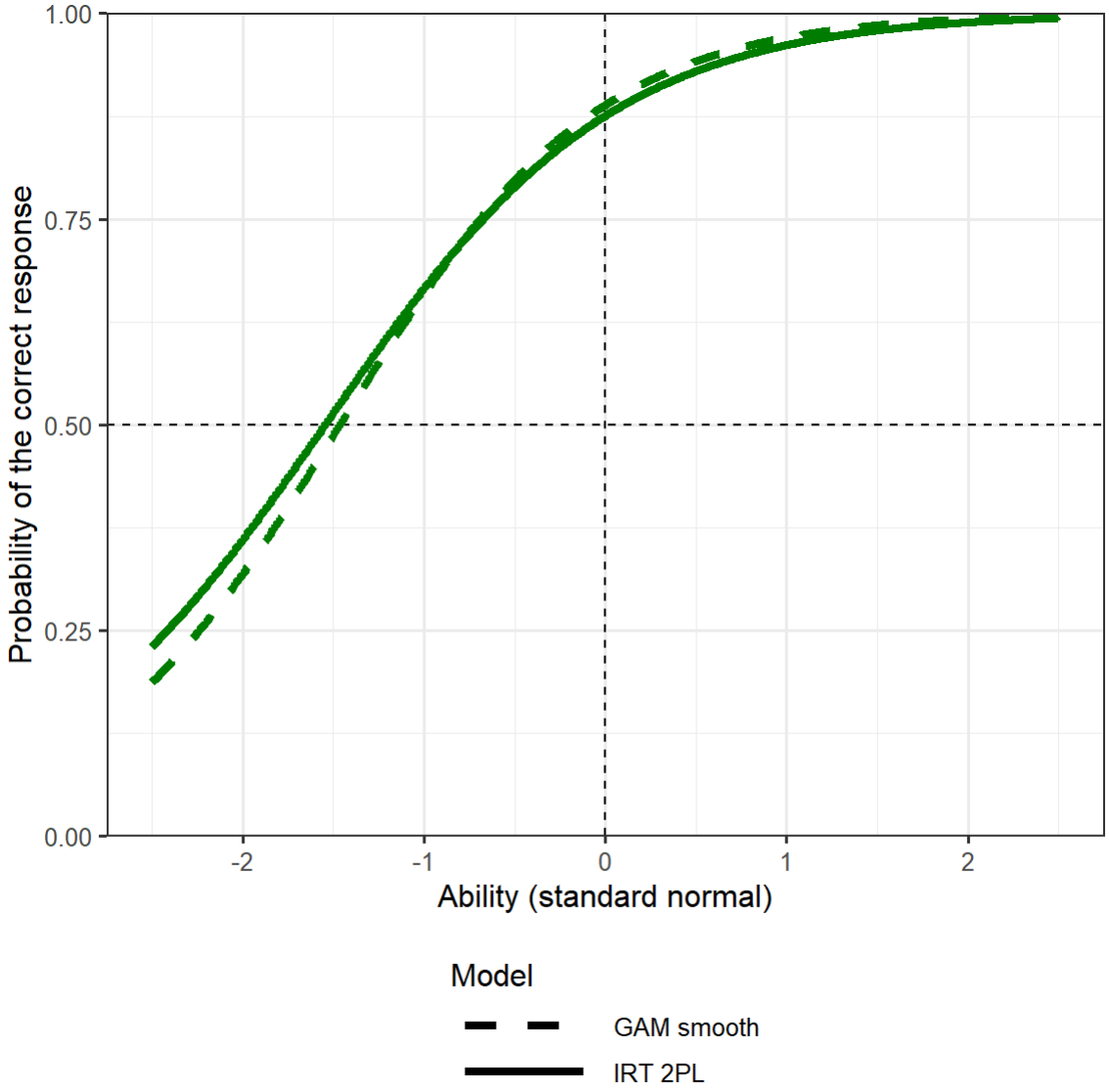
10.2.32.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.3	9.0	15.6	26.6	95.0	2.8	0.8
Middle quartiles	3.8	17.0	24.1	35.4	165.9	3.2	0.6
Upper quartile	9.5	17.7	23.4	31.0	80.7	3.2	0.5
Overall	3.3	15.5	22.7	32.3	165.9	3.1	0.6

10.2.33 S051087 (TIMSS)

10.2.33.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	81.31%	306	TIMSS	MC	0.44	0.40	1.26	−1.54	0.000

10.2.33.2 Gender DIF:

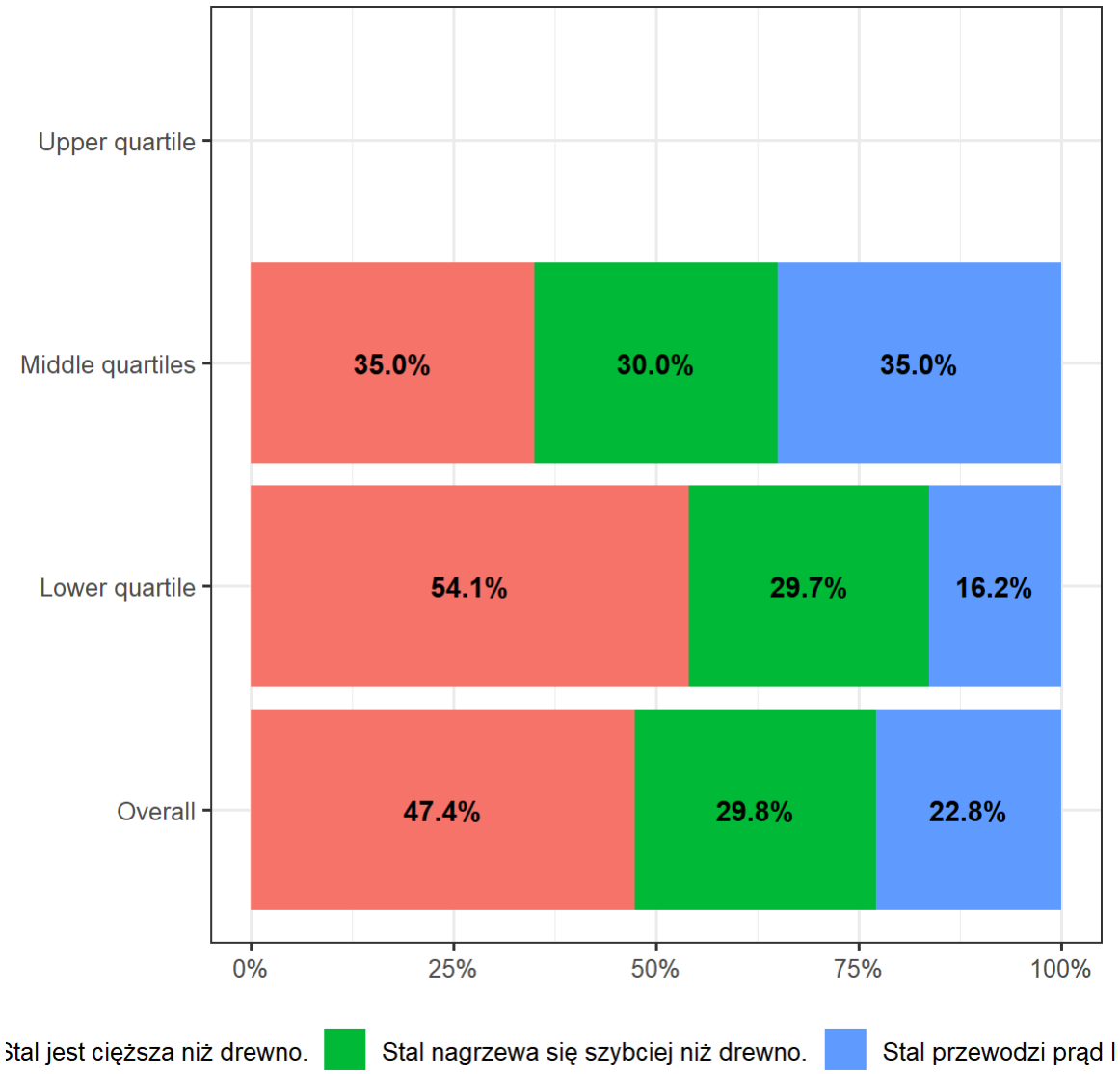
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.54	0.170	1.000	0.013	N/A

10.2.33.3 Frequency of answers:

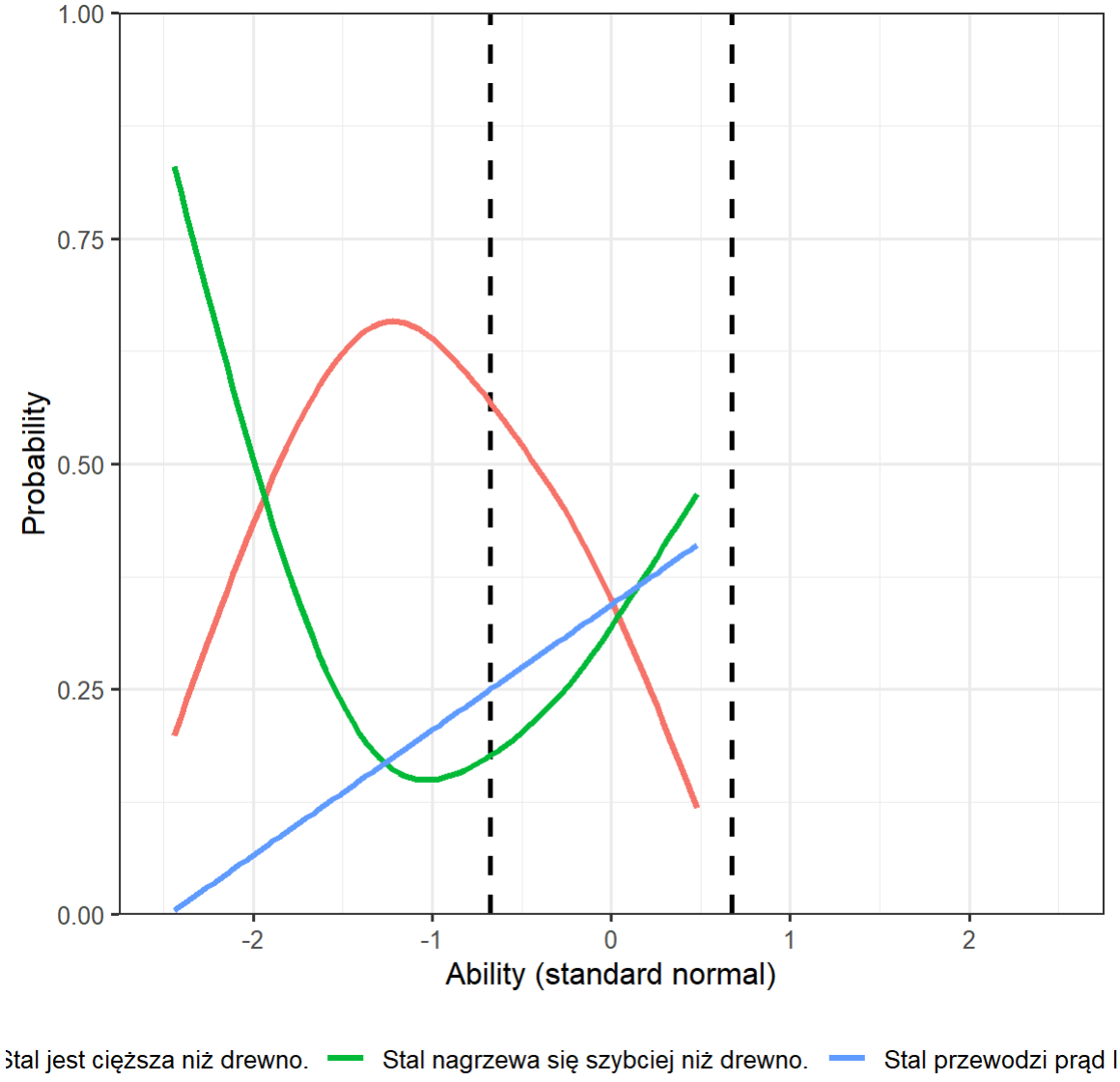
Answer	Frequency	Percent
Stal jest cięższa niż drewno.	27	8.9%
Stal jest mocniejsza niż drewno. (correct)	248	81.3%
Stal nagrzewa się szybciej niż drewno.	17	5.6%
Stal przewodzi prąd lepiej niż drewno.	13	4.3%
Overall	305	100.0%

10.2.33.4 Distractors functioning:

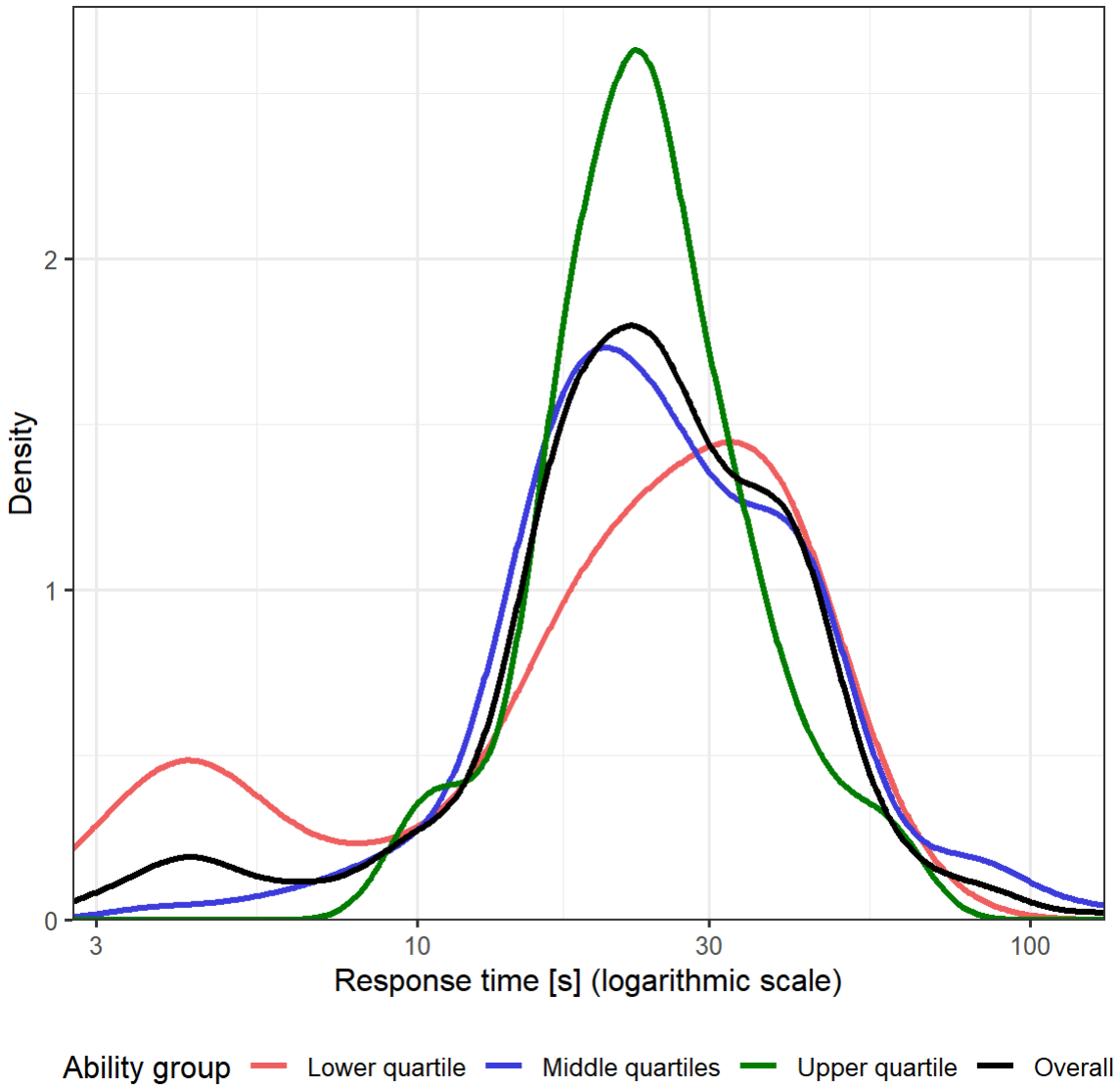
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Stal jest cięższa niż drewno.	20	7	0	27	54.1%	35.0%	0.0%	47.4%
Stal nagrzewa się szybciej niż drewno.	11	6	0	17	29.7%	30.0%	0.0%	29.8%
Stal przewodzi prąd lepiej niż drewno.	6	7	0	13	16.2%	35.0%	0.0%	22.8%
Overall	37	20	0	57	100.0%	100.0%	0.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



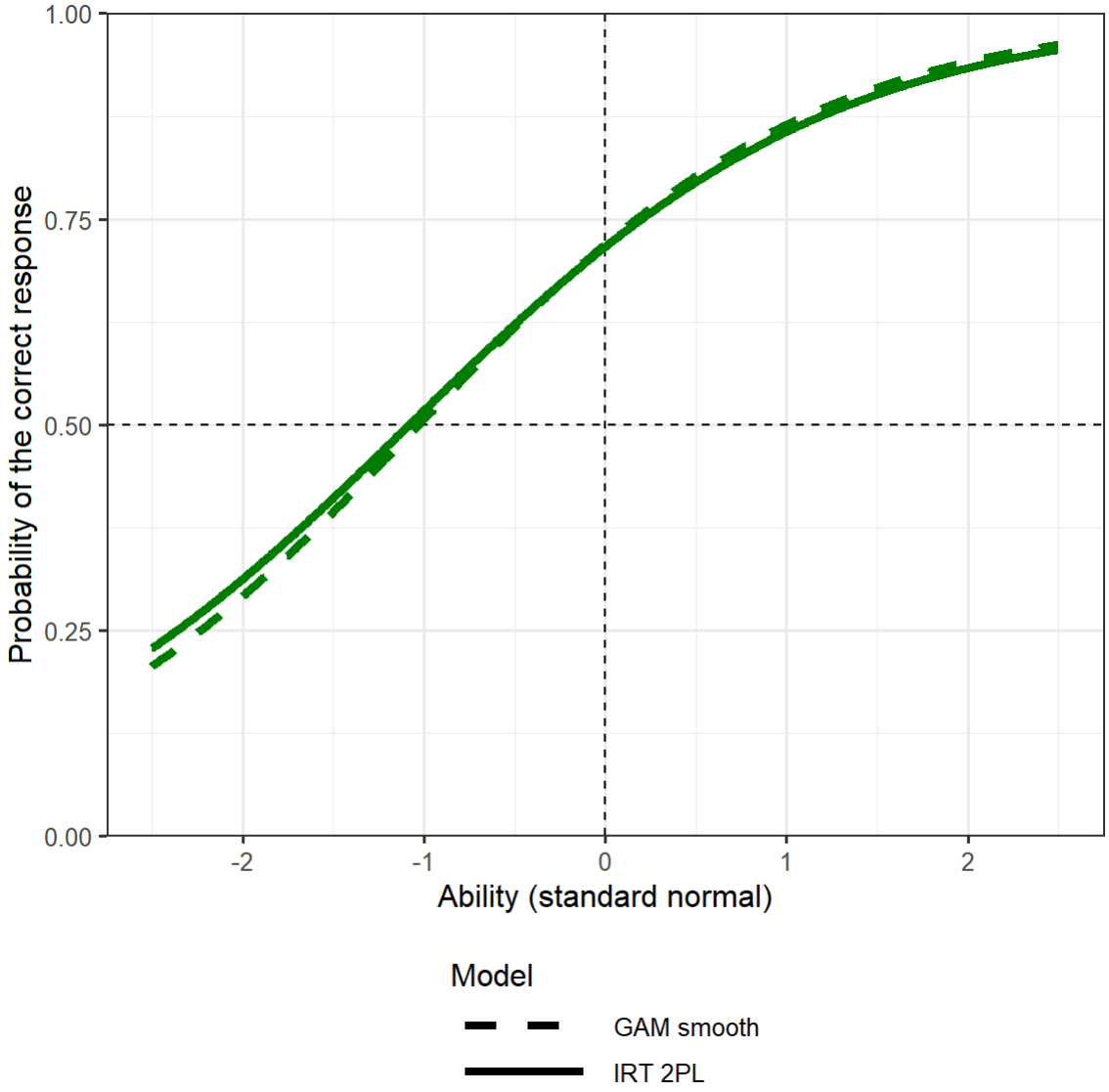
10.2.33.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	14.9	24.1	36.0	65.5	3.0	0.8
Middle quartiles	3.7	17.8	24.1	37.5	132.2	3.2	0.6
Upper quartile	9.6	18.5	23.5	30.2	62.2	3.2	0.4
Overall	2.7	17.3	23.8	35.2	132.2	3.1	0.6

10.2.34 S890010 (LLM)

10.2.34.1 Item fit:



							IRT model including both TIMSS generated items		
					$r_{score,\theta}$				
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	69.28%	319	LLM complex setup	MC	0.36	0.24	0.86	−1.08	0.00

10.2.34.2 Gender DIF:

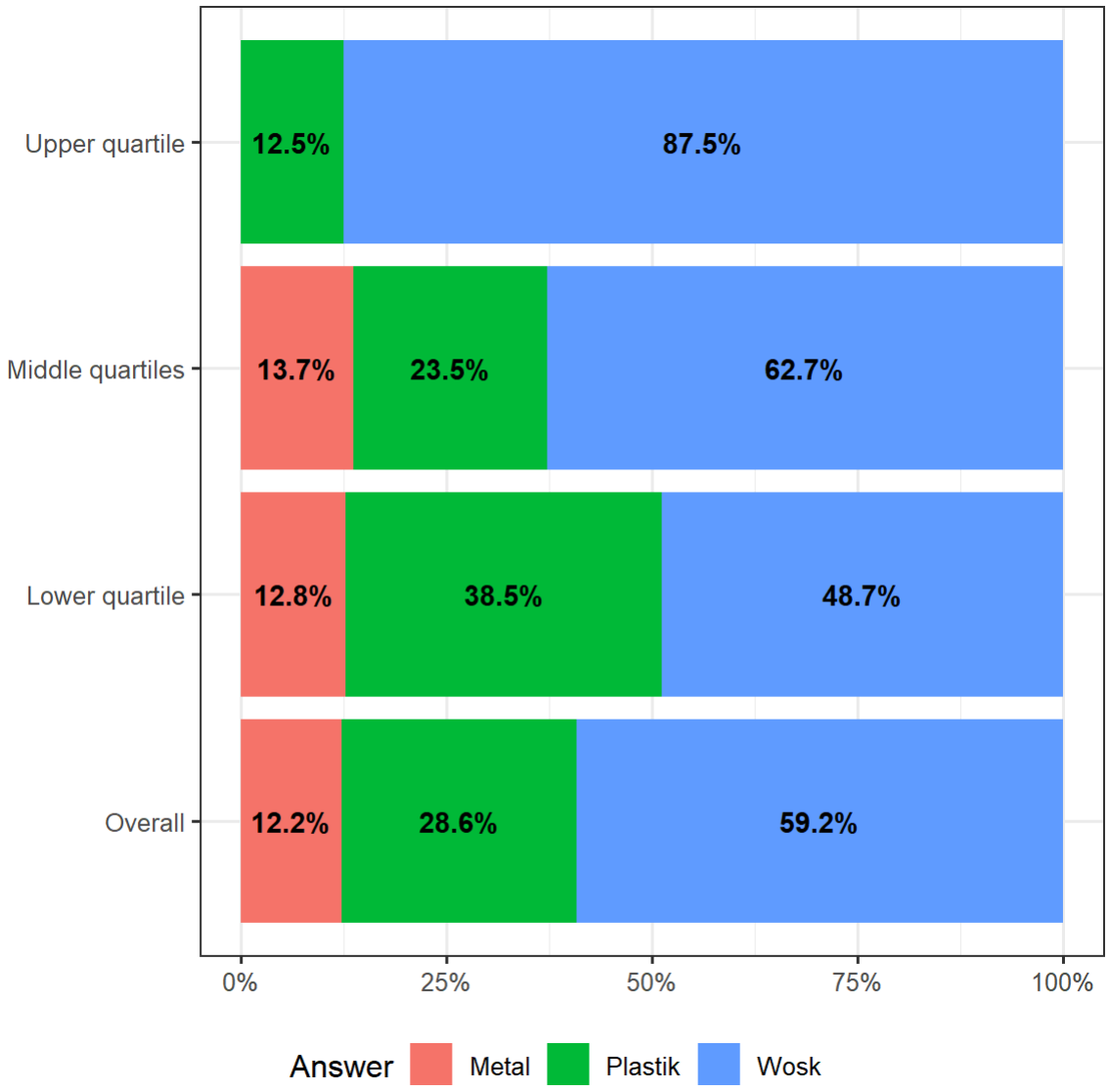
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.48	0.787	1.000	0.001	N/A

10.2.34.3 Frequency of answers:

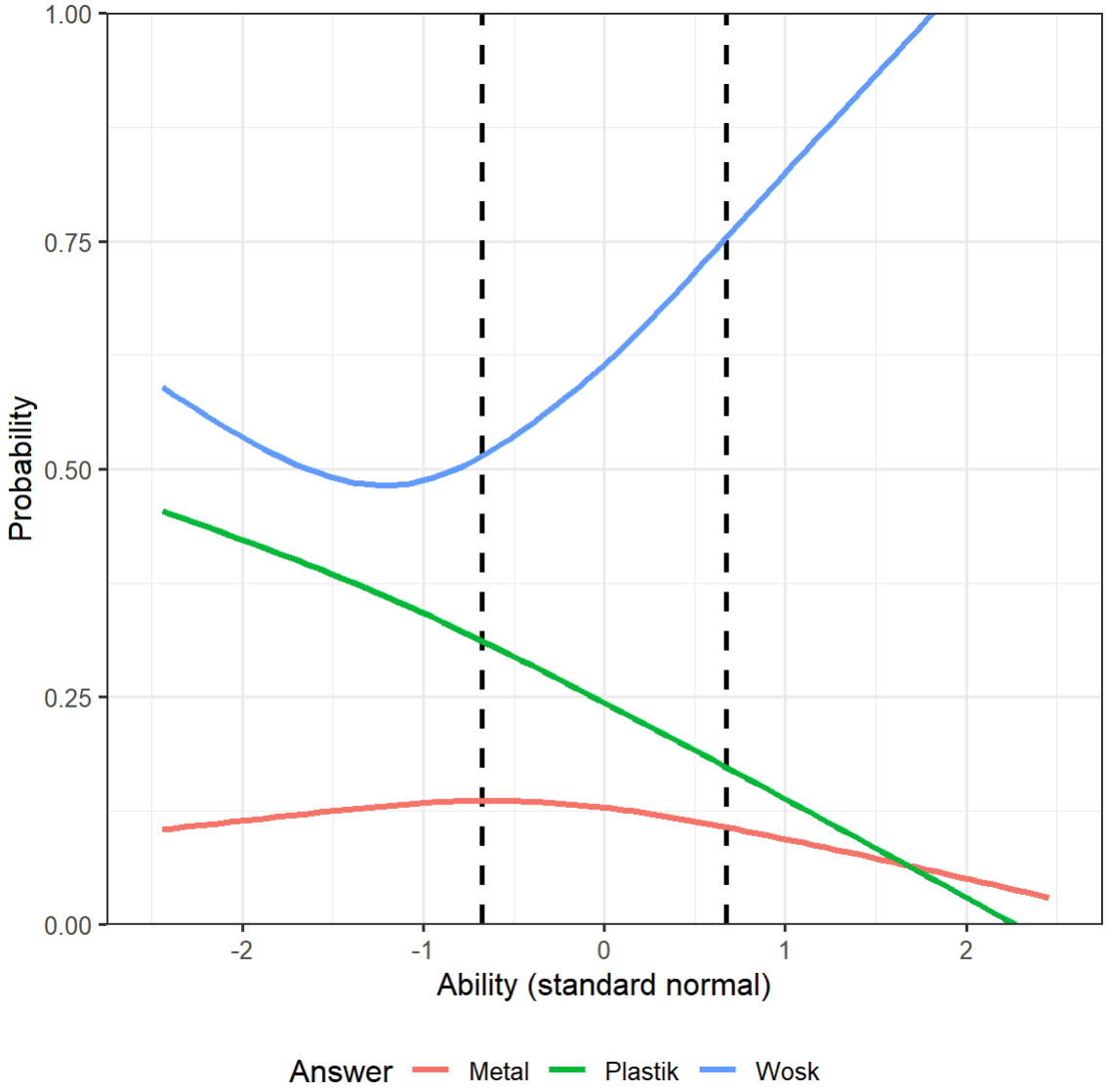
Answer	Frequency	Percent
Lód (correct)	221	69.3%
Metal	12	3.8%
Plastik	28	8.8%
Wosk	58	18.2%
Overall	319	100.0%

10.2.34.4 Distractors functioning:

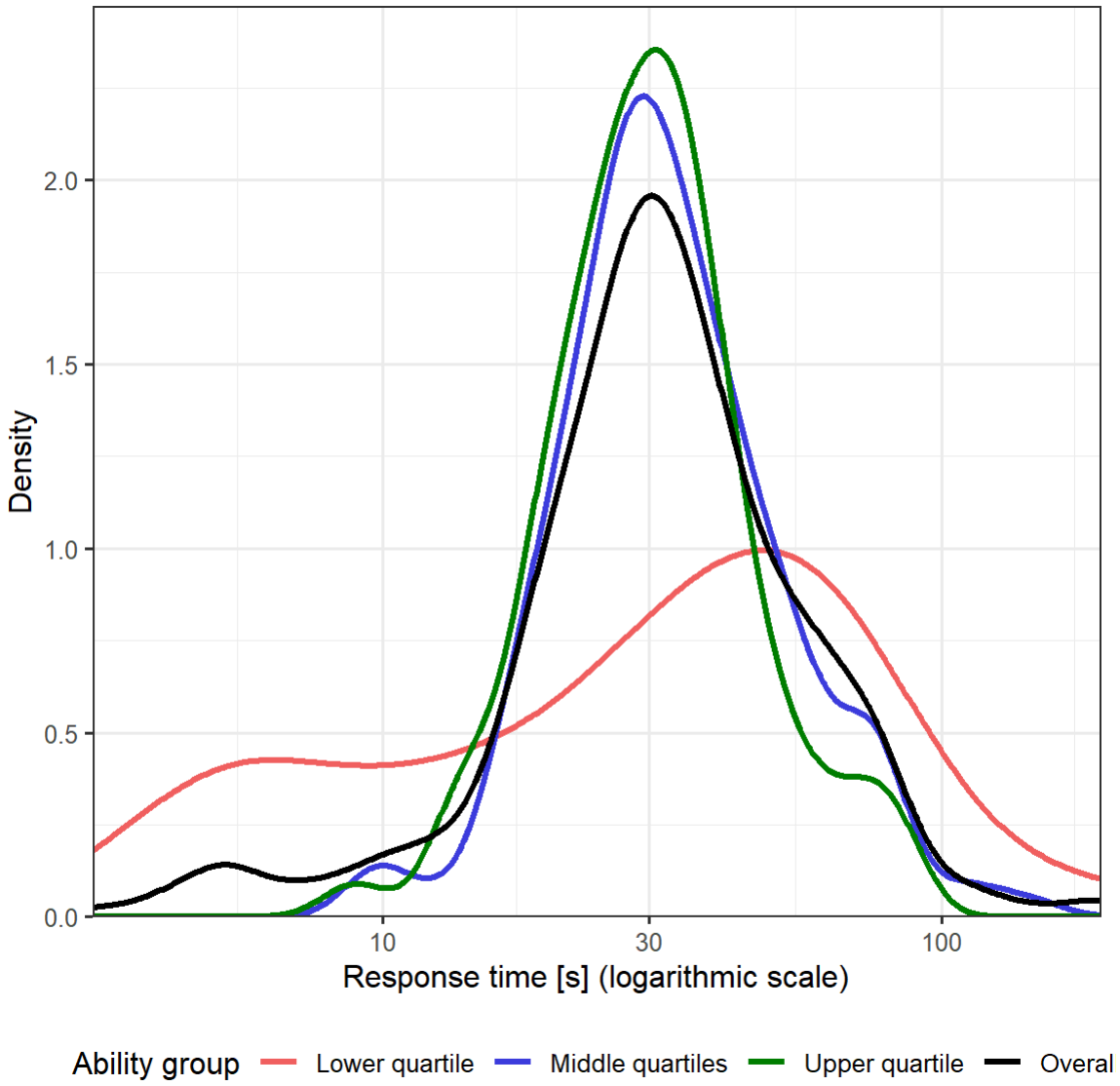
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Metal	5	7	0	12	12.8%	13.7%	0.0%	12.2%
Plastik	15	12	1	28	38.5%	23.5%	12.5%	28.6%
Wosk	19	32	7	58	48.7%	62.7%	87.5%	59.2%
Overall	39	51	8	98	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



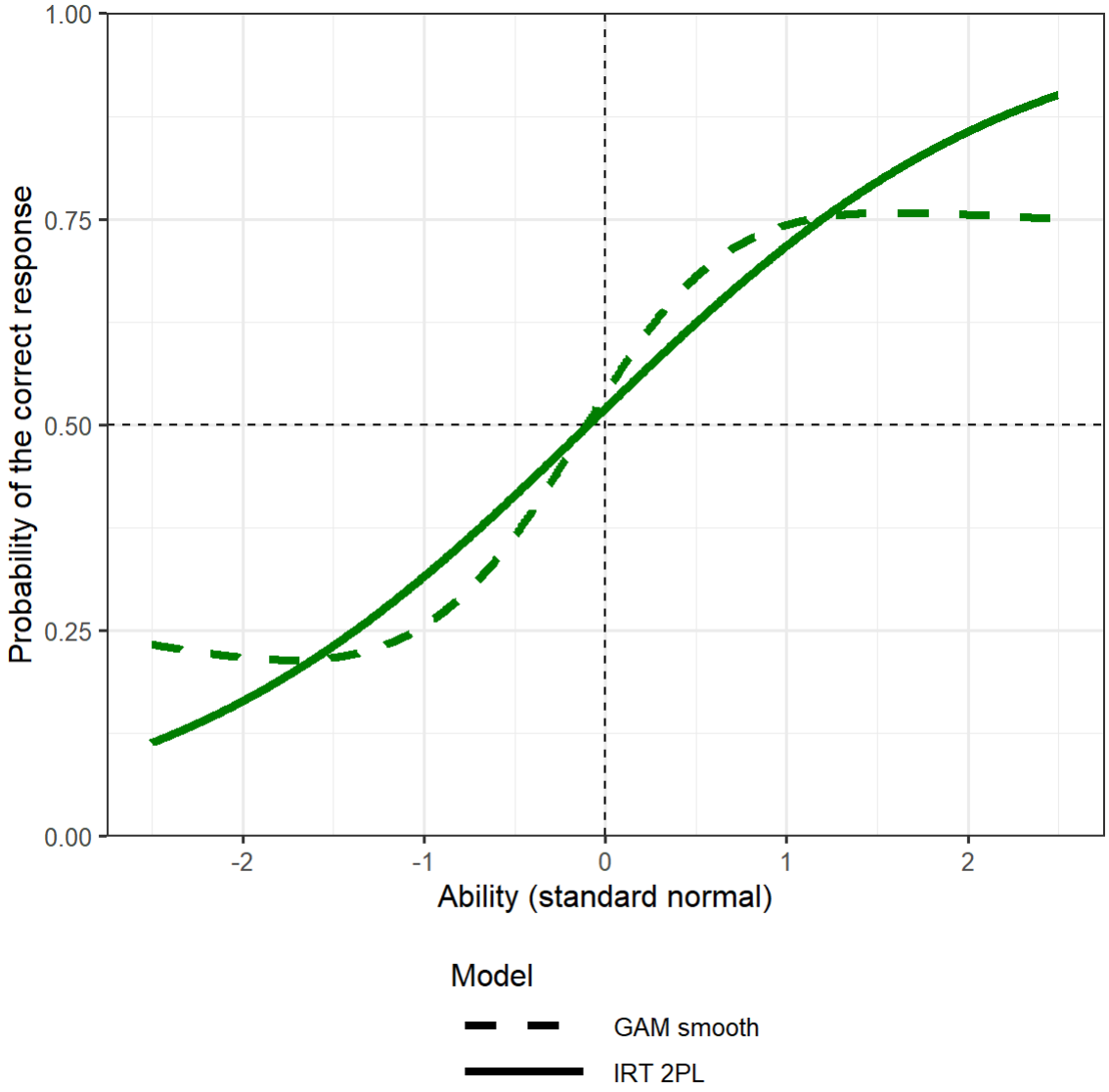
10.2.34.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	12.0	32.2	58.5	192.7	3.3	1.0
Middle quartiles	9.6	24.7	30.7	42.7	145.2	3.5	0.5
Upper quartile	8.9	23.4	30.3	37.5	82.8	3.4	0.4
Overall	3.0	22.3	30.7	43.1	192.7	3.4	0.6

10.2.35 SP71072 (TIMSS)

10.2.35.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	52.72%	294	TIMSS	MC	0.36	0.42	0.85	−0.09	0.017

10.2.35.2 Gender DIF:

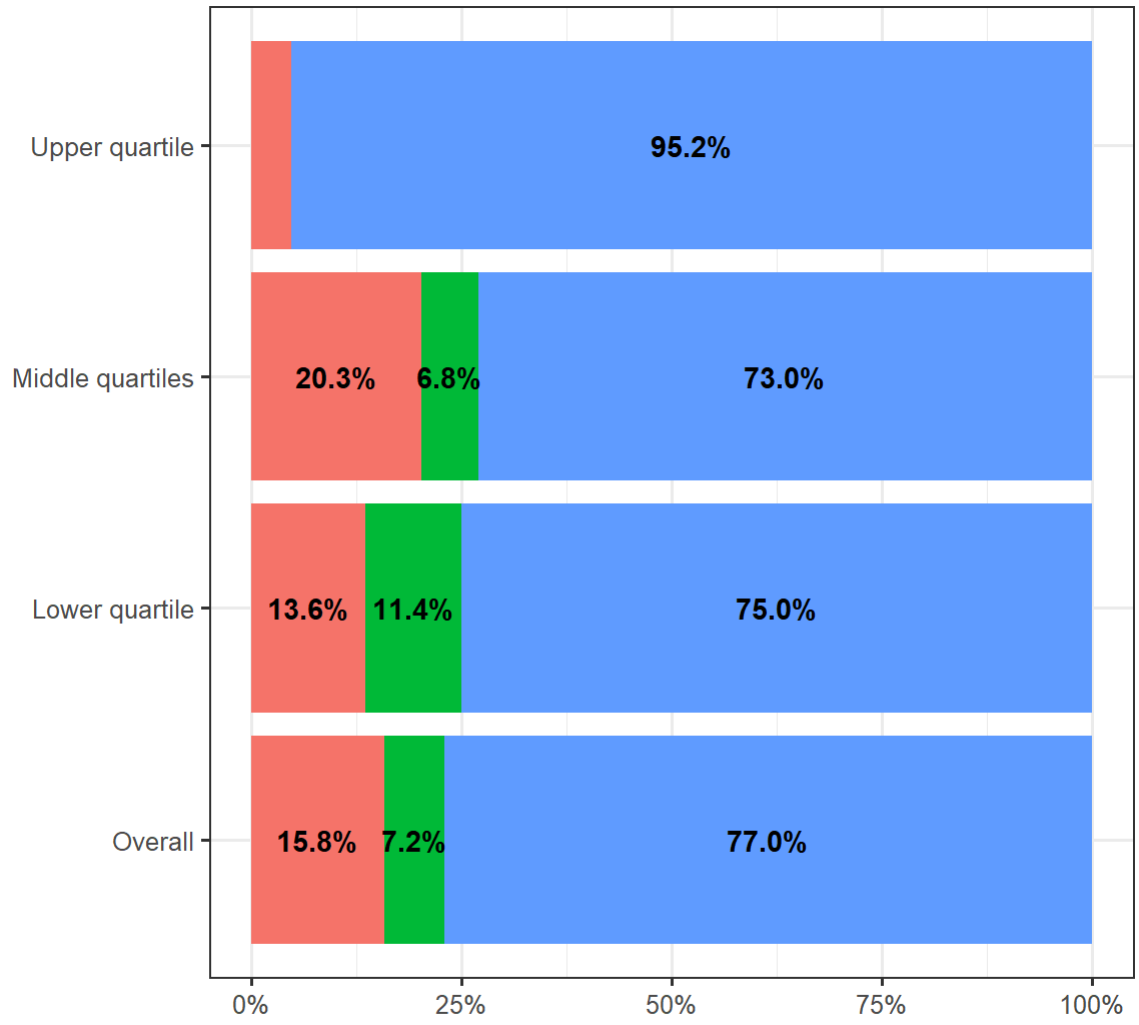
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
5.03	0.081	1.000	0.014	N/A

10.2.35.3 Frequency of answers:

Answer	Frequency	Percent
Muchołówka pobiera składniki odżywcze od owadów, a inne rośliny tego nie robią. (correct)	155	52.7%
Muchołówka pobiera wodę z owadów, a inne rośliny tego nie robią.	22	7.5%
Muchołówka pomaga w rozmnażaniu się owadów, a inne rośliny tego nie robią.	10	3.4%
Muchołówka przyciąga owady, a inne rośliny tego nie robią.	107	36.4%
Overall	294	100.0%

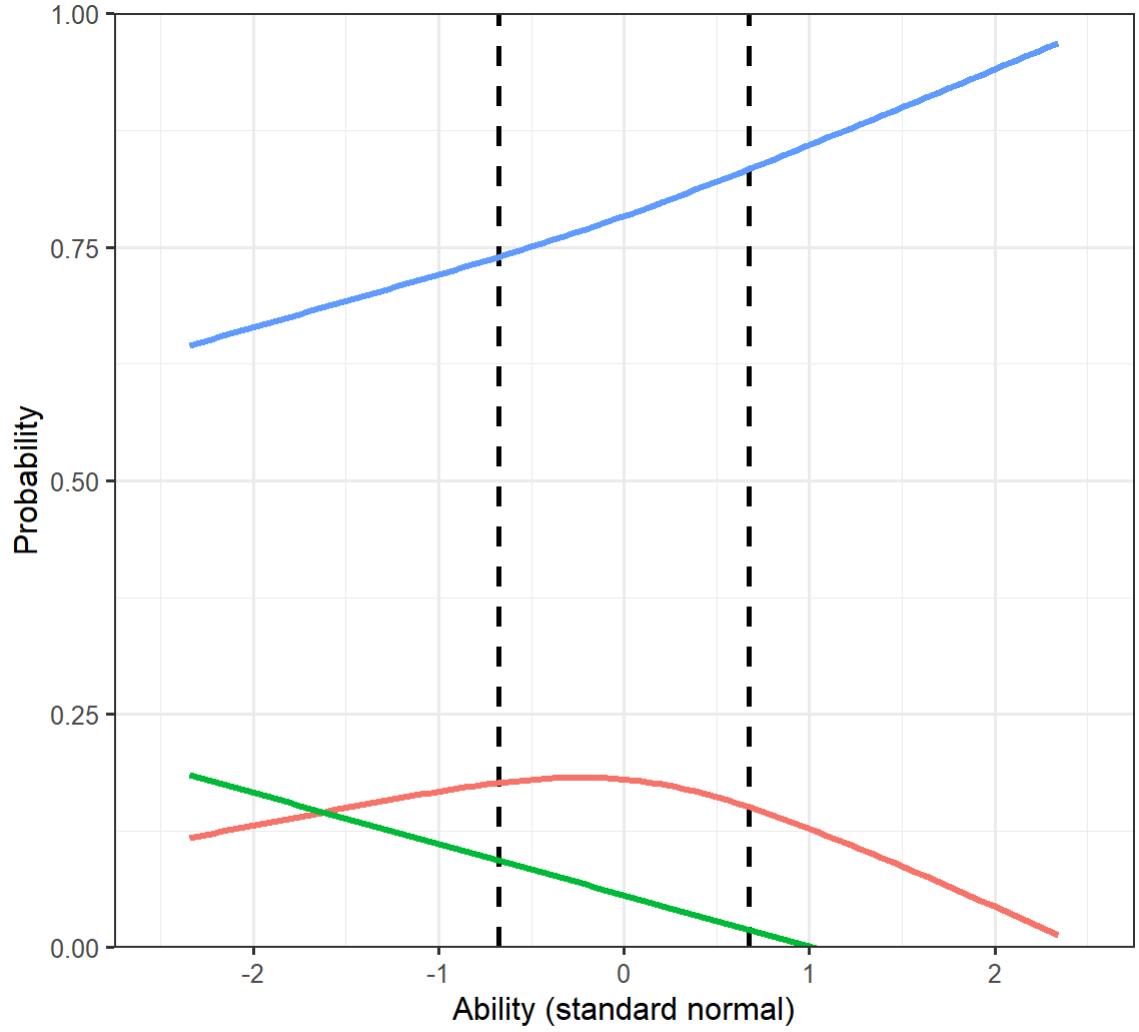
10.2.35.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Muchołówka pobiera wodę z owadów, a inne rośliny tego nie robią.	6	15	1	22	13.6%	20.3%	4.8%	15.8%
Muchołówka pomaga w rozmnażaniu się owadów, a inne rośliny tego nie robią.	5	5	0	10	11.4%	6.8%	0.0%	7.2%
Muchołówka przyciąga owady, a inne rośliny tego nie robią.	33	54	20	107	75.0%	73.0%	95.2%	77.0%
Overall	44	74	21	139	100.0%	100.0%	100.0%	100.0%



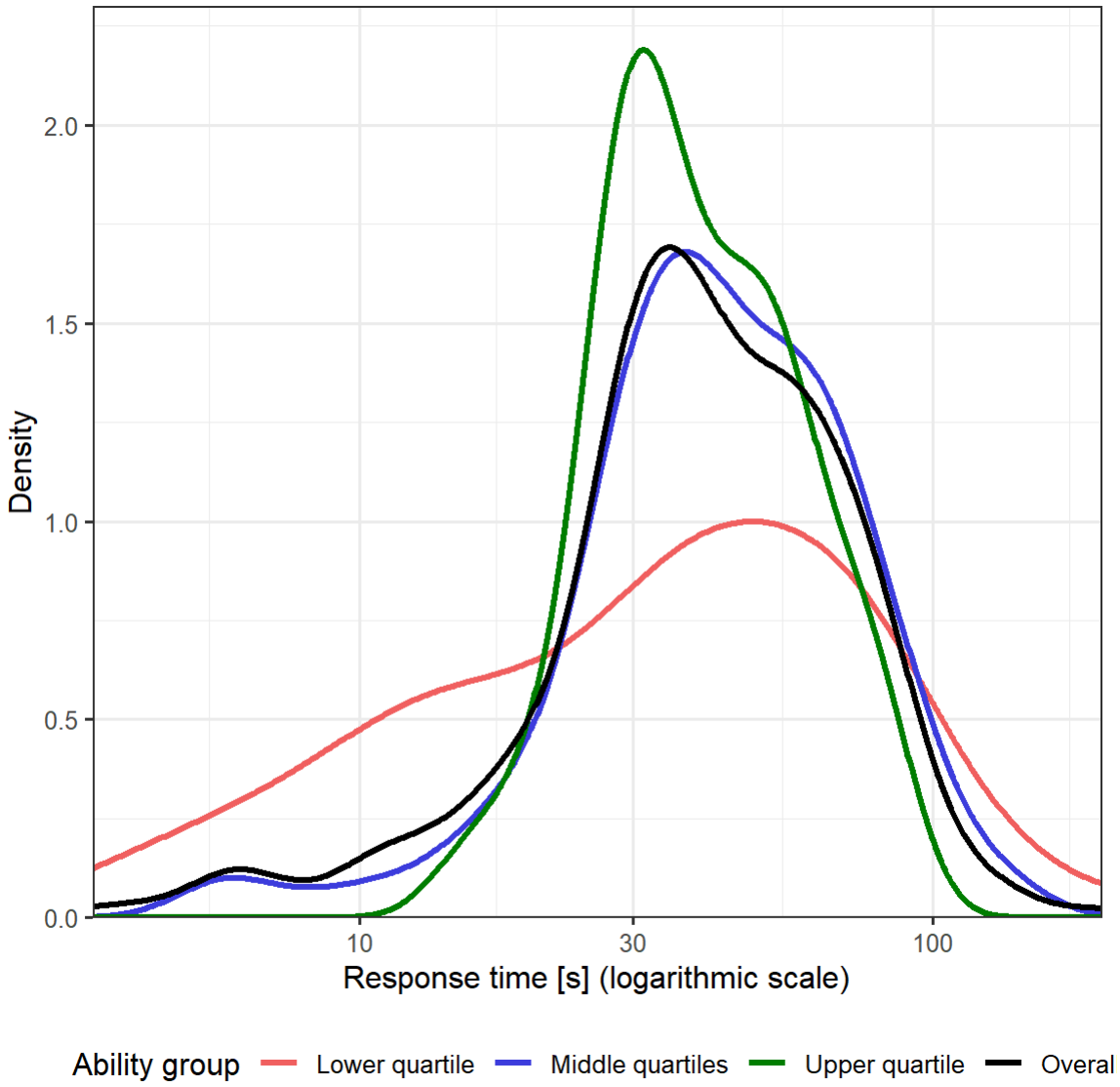
y tego nie robią. Muchołówka pomaga w rozmnażaniu się owadów, a inne rośliny tego nie robia

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



y tego nie robią. Muchołówka pomaga w rozmnażaniu się owadów, a inne rośliny tego nie robia

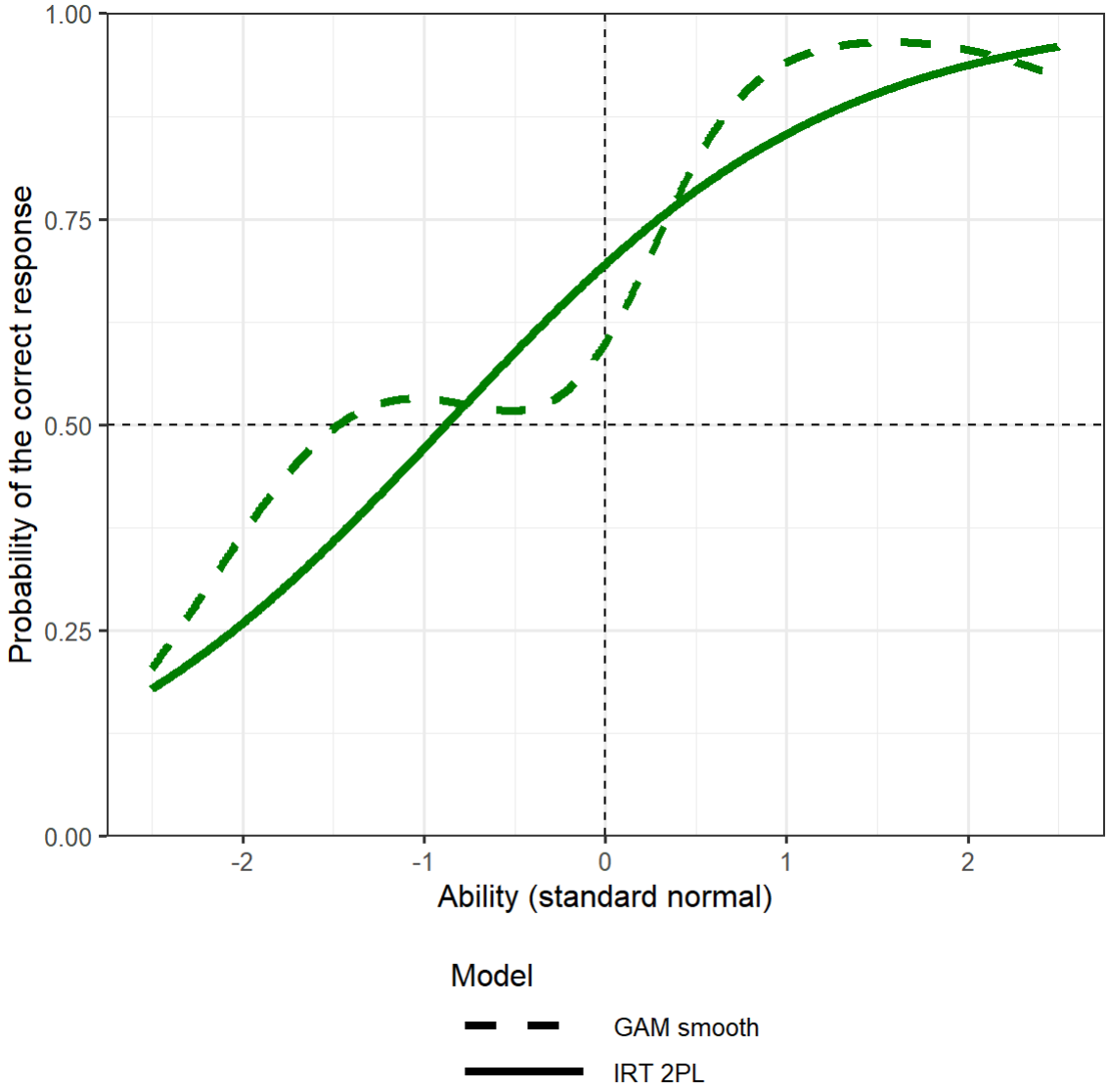
10.2.35.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.4	15.8	36.1	61.5	196.7	3.4	0.9
Middle quartiles	5.2	30.1	41.0	60.8	144.1	3.7	0.6
Upper quartile	14.8	29.0	37.6	52.3	86.7	3.7	0.4
Overall	3.4	28.6	39.9	58.9	196.7	3.6	0.6

10.2.36 S041052 (TIMSS)

10.2.36.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	66.12%	305	TIMSS	MC	0.36	0.39	0.94	−0.88	0.037

10.2.36.2 Gender DIF:

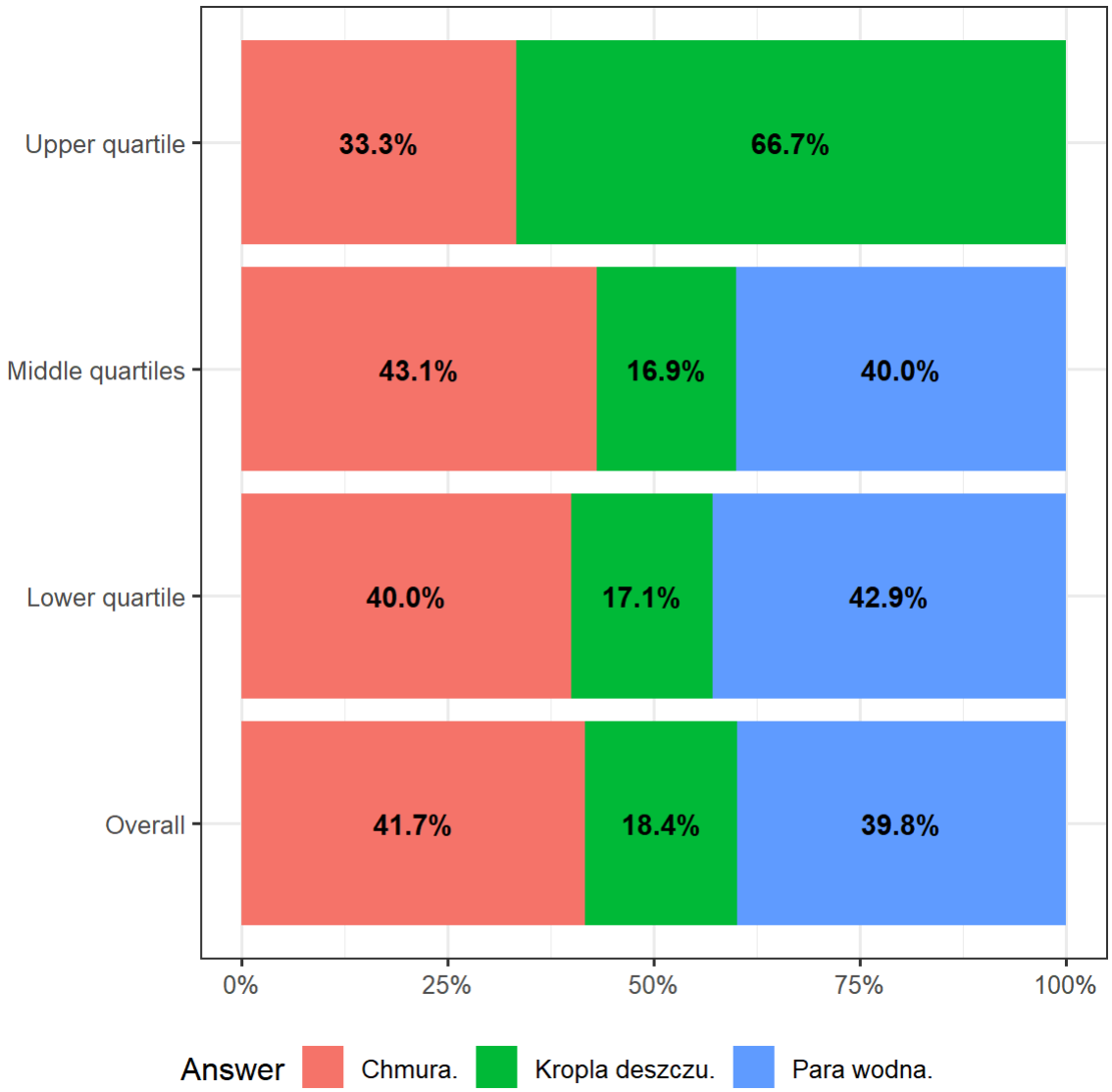
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
4.26	0.119	1.000	0.012	N/A

10.2.36.3 Frequency of answers:

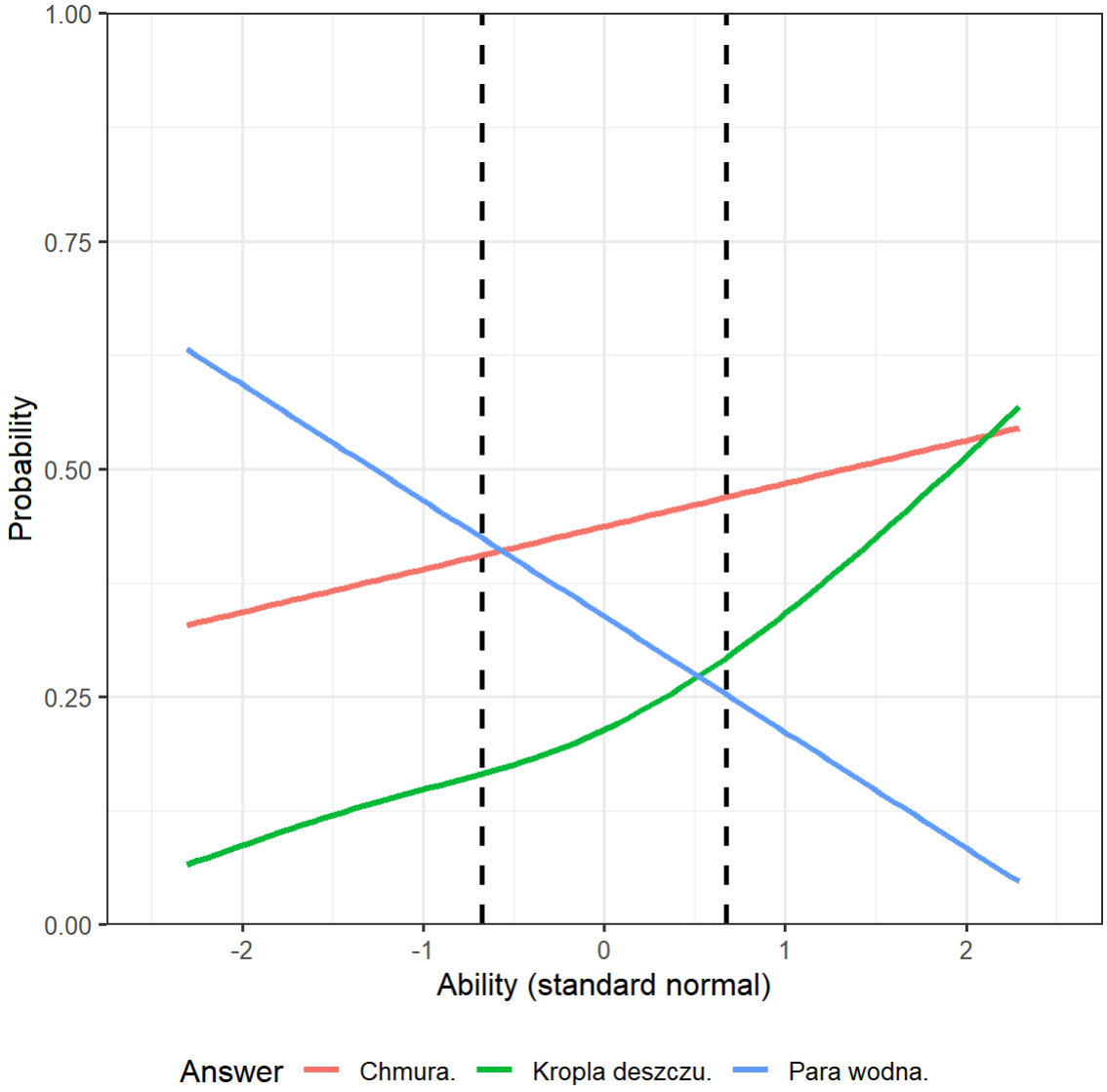
Answer	Frequency	Percent
Chmura.	43	14.1%
Kostka lodu. (correct)	201	66.1%
Kropla deszczu.	19	6.2%
Para wodna.	41	13.5%
Overall	304	100.0%

10.2.36.4 Distractors functioning:

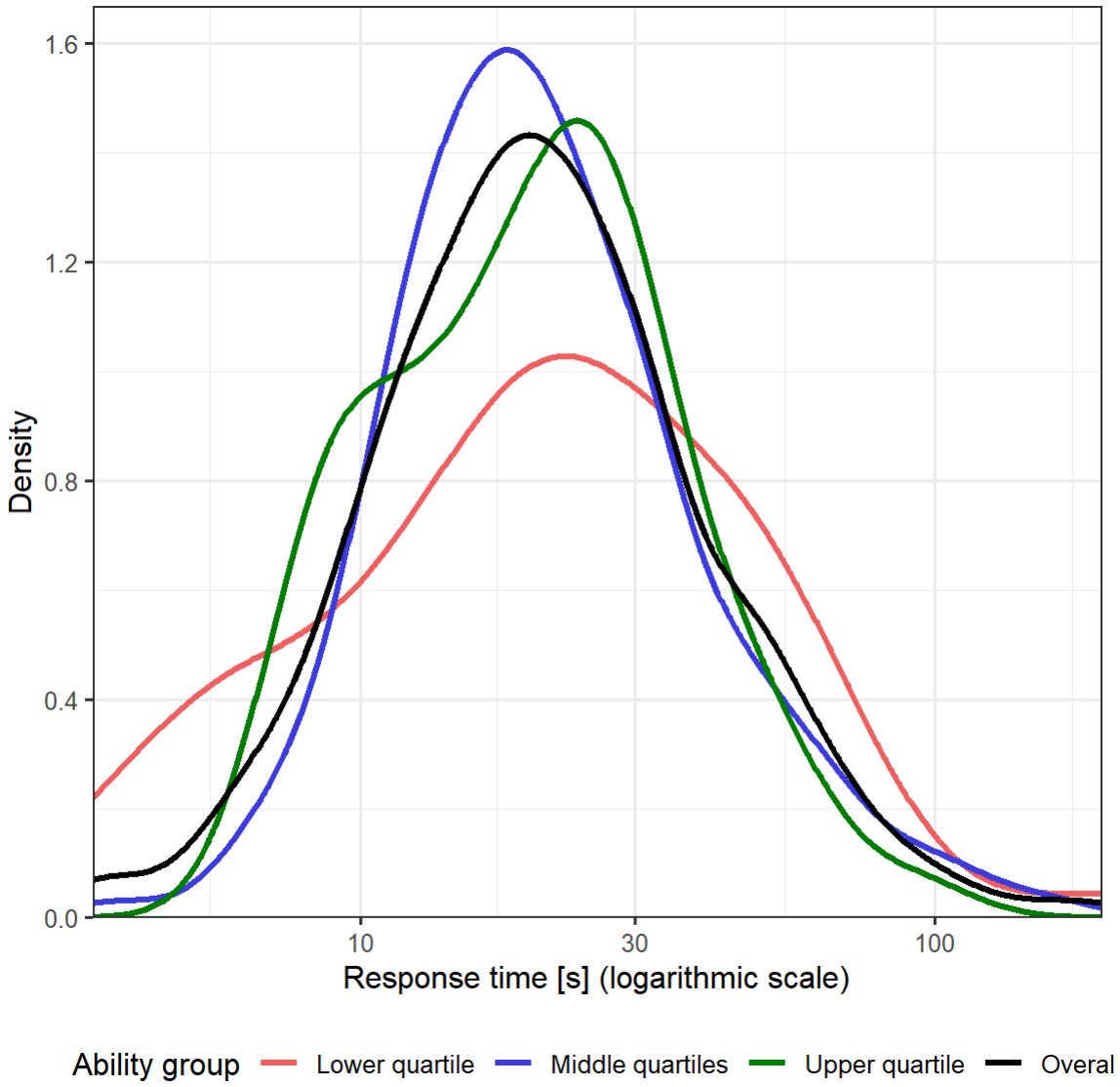
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Chmura.	14	28	1	43	40.0%	43.1%	33.3%	41.7%
Kropla deszczu.	6	11	2	19	17.1%	16.9%	66.7%	18.4%
Para wodna.	15	26	0	41	42.9%	40.0%	0.0%	39.8%
Overall	35	65	3	103	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



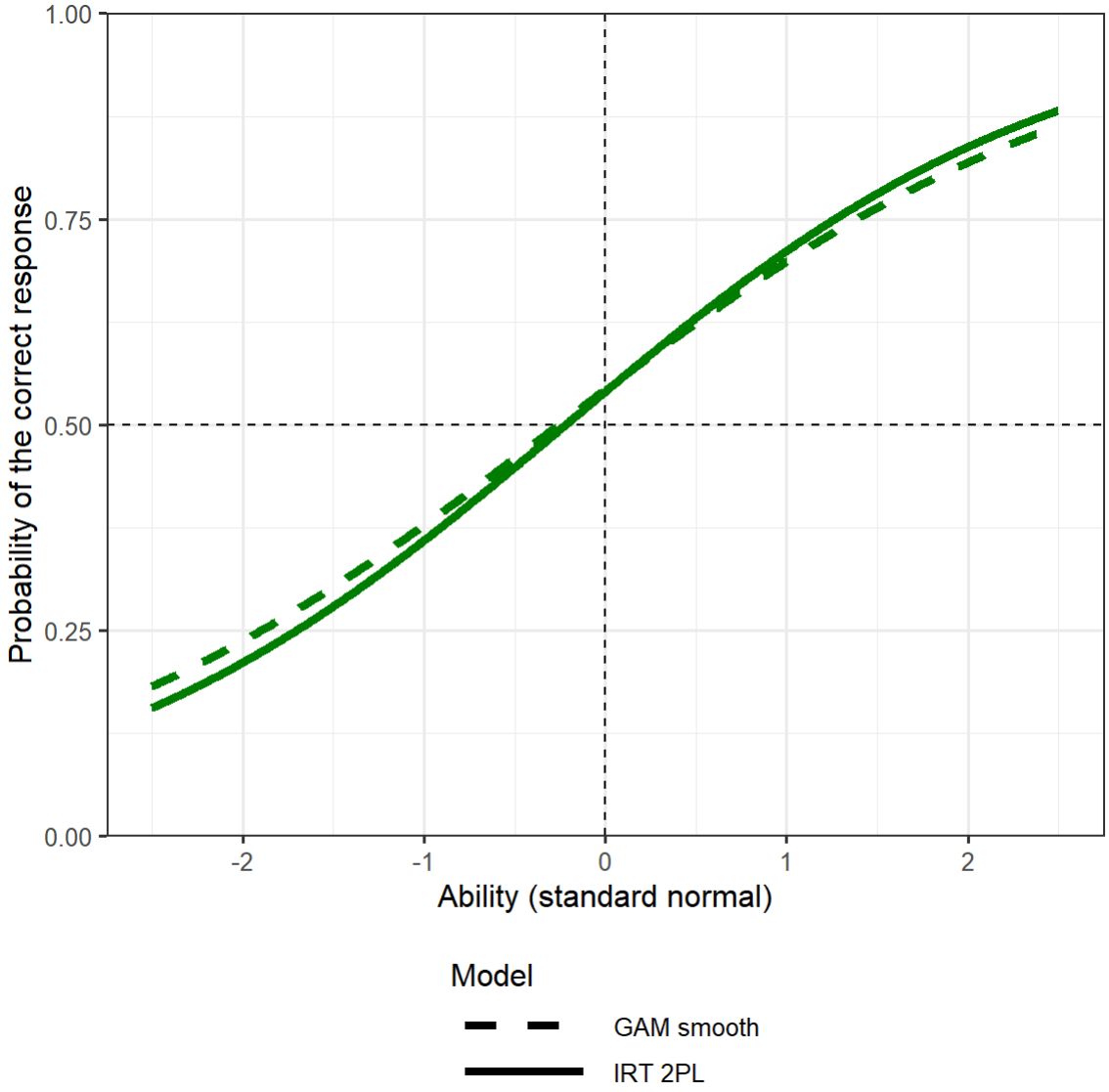
10.2.36.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.4	11.2	21.2	36.0	194.6	3.0	0.9
Middle quartiles	3.7	13.6	19.8	29.9	160.9	3.0	0.6
Upper quartile	7.2	12.7	21.6	29.0	92.5	3.0	0.6
Overall	3.4	13.1	20.2	30.5	194.6	3.0	0.7

10.2.37 S041070 (TIMSS)

10.2.37.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	52.79%	306	TIMSS	MC	0.30	0.37	0.74	−0.22	0.008

10.2.37.2 Gender DIF:

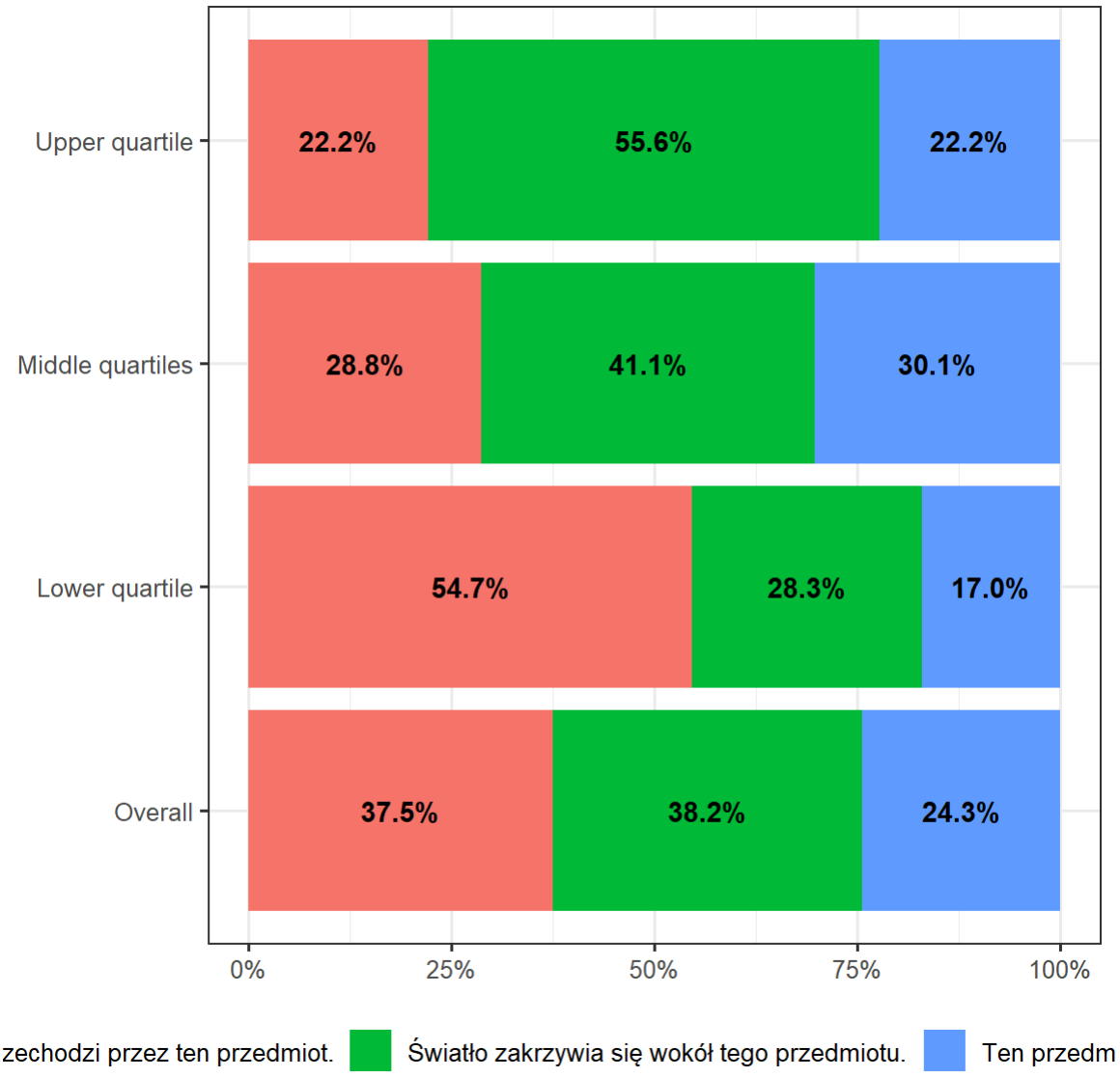
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.42	0.181	1.000	0.009	N/A

10.2.37.3 Frequency of answers:

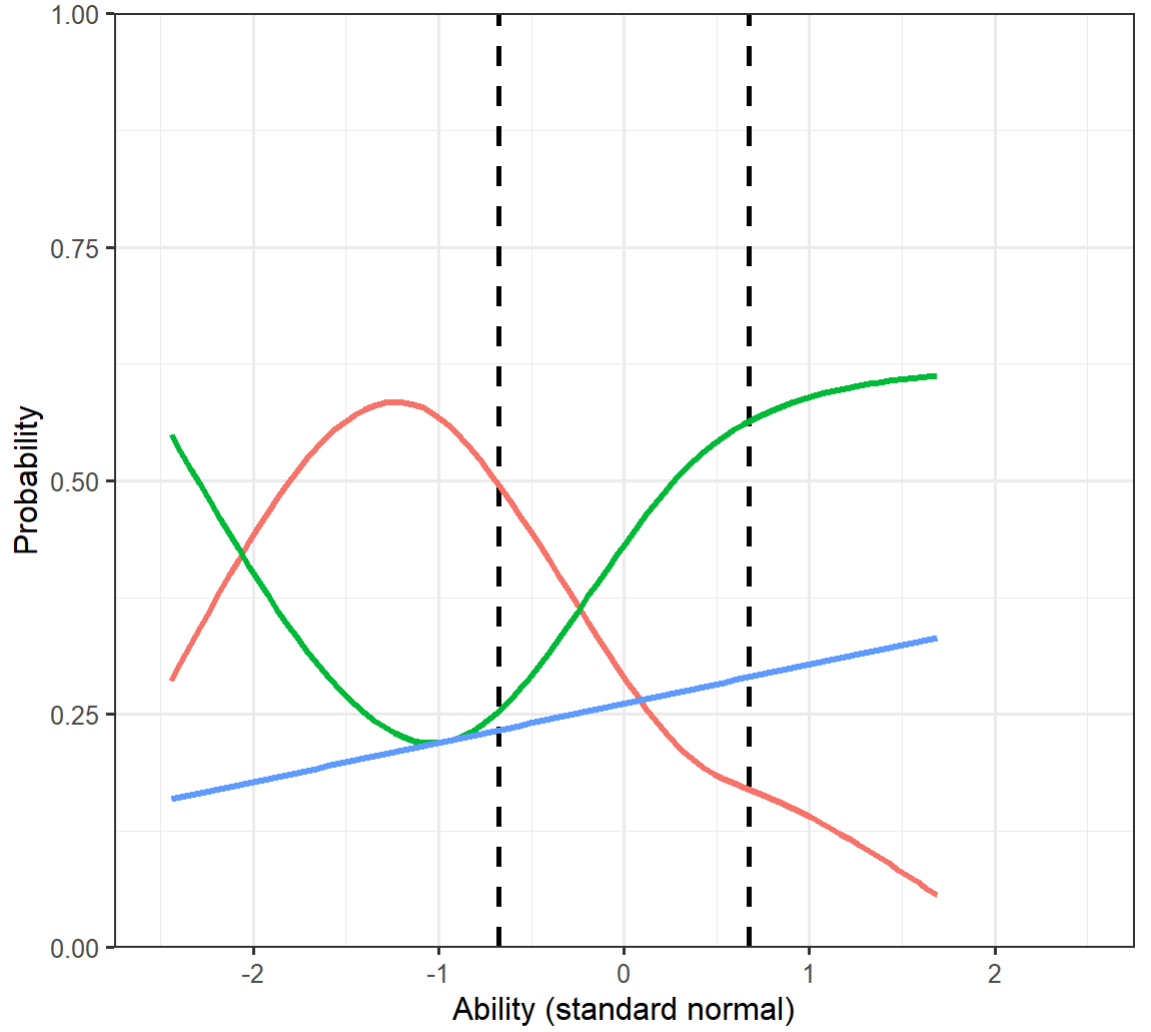
Answer	Frequency	Percent
Ten przedmiot odchyla na bok światło.	35	11.5%
Ten przedmiot zatrzymuje światło. (correct)	161	52.8%
Światło przechodzi przez ten przedmiot.	54	17.7%
Światło zakrzywia się wokół tego przedmiotu.	55	18.0%
Overall	305	100.0%

10.2.37.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Ten przedmiot odchyla na bok światło.	9	22	4	35	17.0%	30.1%	22.2%	24.3%
Światło przechodzi przez ten przedmiot.	29	21	4	54	54.7%	28.8%	22.2%	37.5%
Światło zakrzywia się wokół tego przedmiotu.	15	30	10	55	28.3%	41.1%	55.6%	38.2%
Overall	53	73	18	144	100.0%	100.0%	100.0%	100.0%

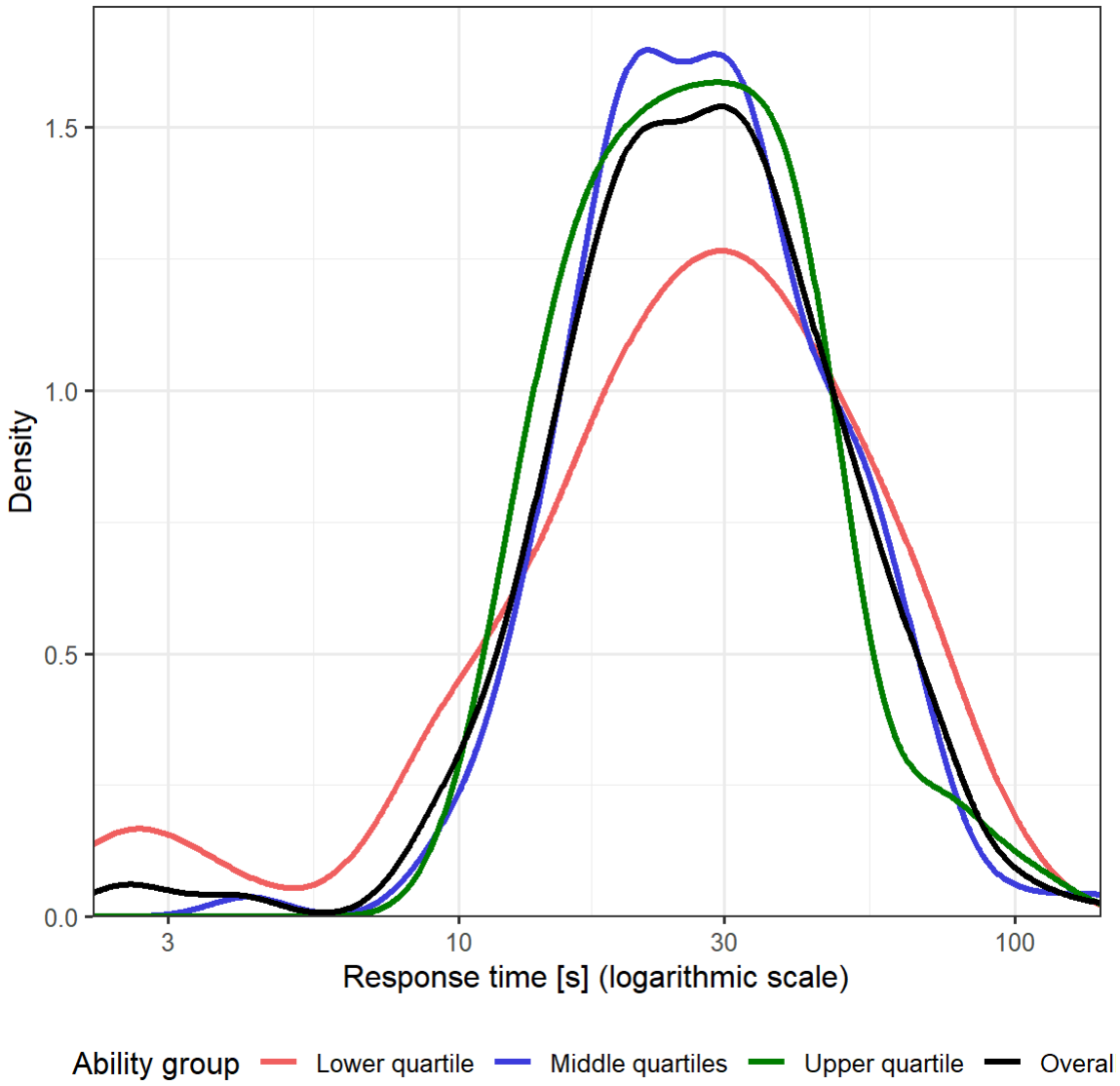



```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



zechodzi przez ten przedmiot. Światło zakrzywia się wokół tego przedmiotu. Ten przedm

10.2.37.5 Answering time:

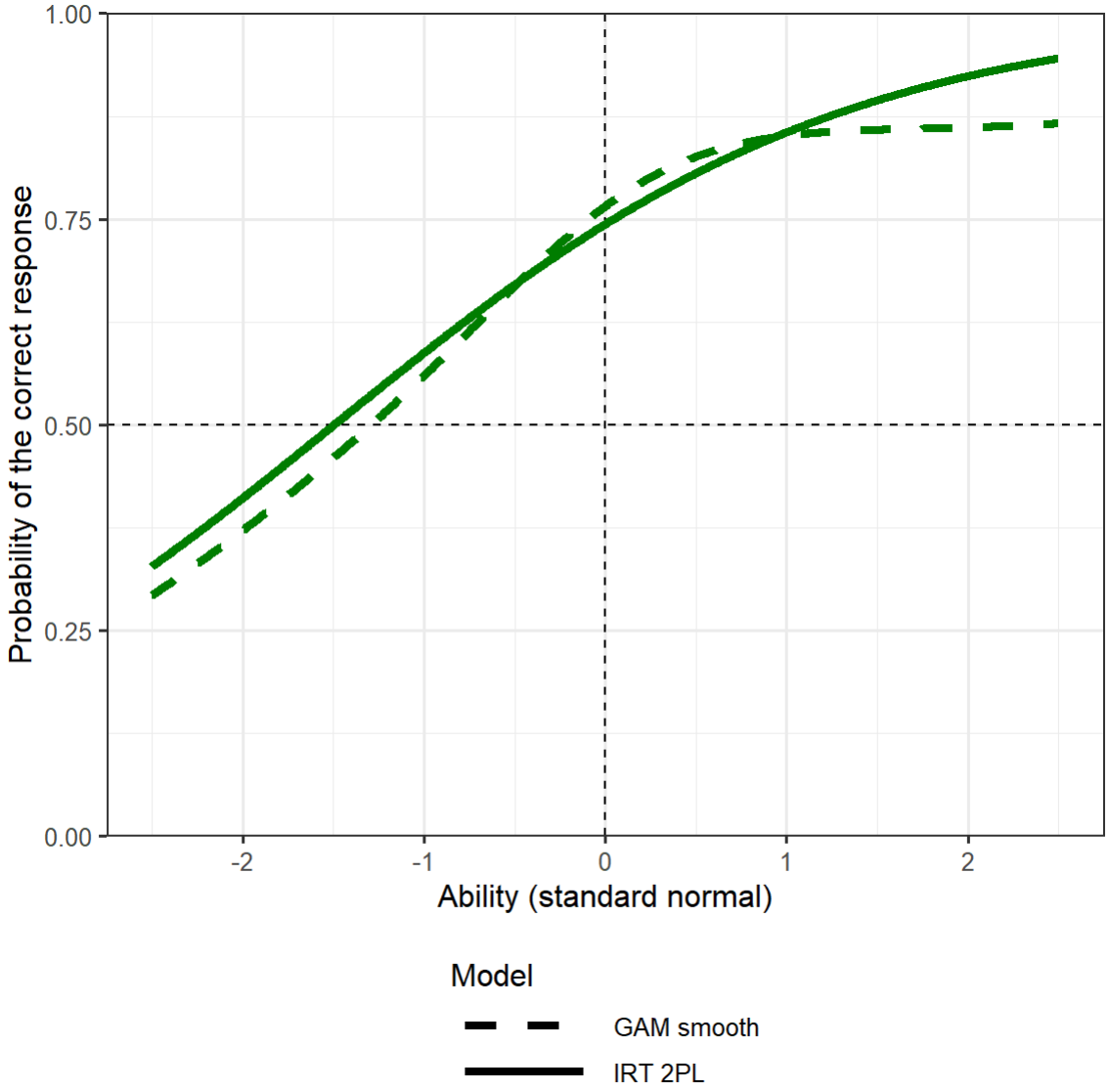


	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.2	16.8	28.0	41.9	98.3	3.2	0.8
Middle quartiles	4.2	19.8	27.1	38.3	142.6	3.3	0.5
Upper quartile	10.7	18.4	25.8	37.6	108.1	3.3	0.5
Overall	2.2	18.6	27.1	39.2	142.6	3.3	0.6

10.2.38 S100016 (TIMSS)

10.2.38.1 Item fit:

```
## Warning in newton(lsp = lsp, X = G$X, y = G$y, Eb = G$Eb, UrS = G$UrS, L = G$L,
## : Fitting terminated with step failure - check results carefully
## Warning in newton(lsp = lsp, X = G$X, y = G$y, Eb = G$Eb, UrS = G$UrS, L = G$L,
## : Fitting terminated with step failure - check results carefully
```



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	72.04%	305	LLM simple setup	MC	0.28	0.18	0.71	−1.49	0.015

10.2.38.2 Gender DIF:

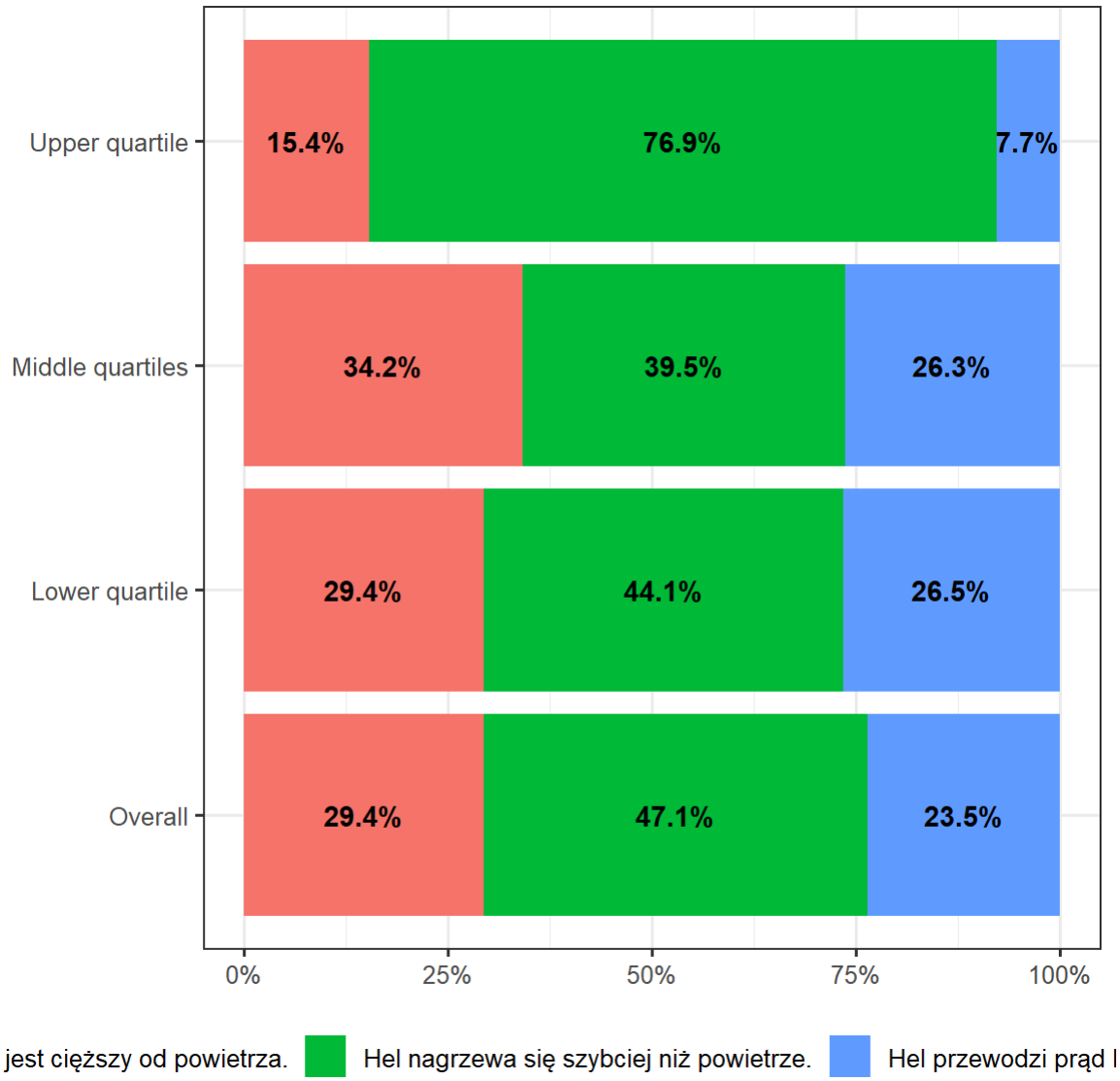
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.25	0.534	1.000	0.004	N/A

10.2.38.3 Frequency of answers:

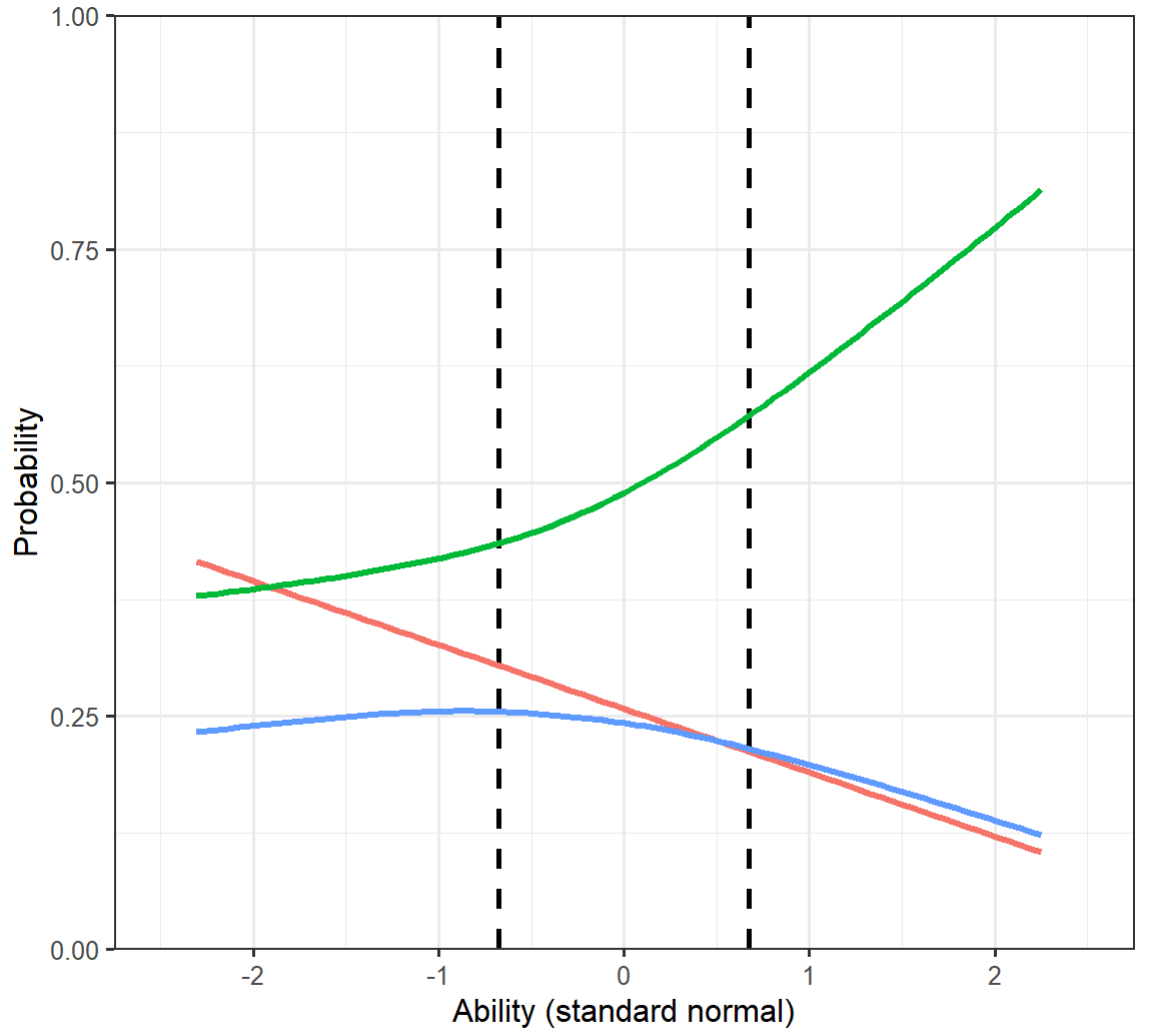
Answer	Frequency	Percent
Hel jest cięższy od powietrza.	25	8.2%
Hel jest lżejszy od powietrza. (correct)	219	72.0%
Hel nagrzewa się szybciej niż powietrze.	40	13.2%
Hel przewodzi prąd lepiej niż powietrze.	20	6.6%
Overall	304	100.0%

10.2.38.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Hel jest cięższy od powietrza.	10	13	2	25	29.4%	34.2%	15.4%	29.4%
Hel nagrzewa się szybciej niż powietrze.	15	15	10	40	44.1%	39.5%	76.9%	47.1%
Hel przewodzi prąd lepiej niż powietrze.	9	10	1	20	26.5%	26.3%	7.7%	23.5%
Overall	34	38	13	85	100.0%	100.0%	100.0%	100.0%

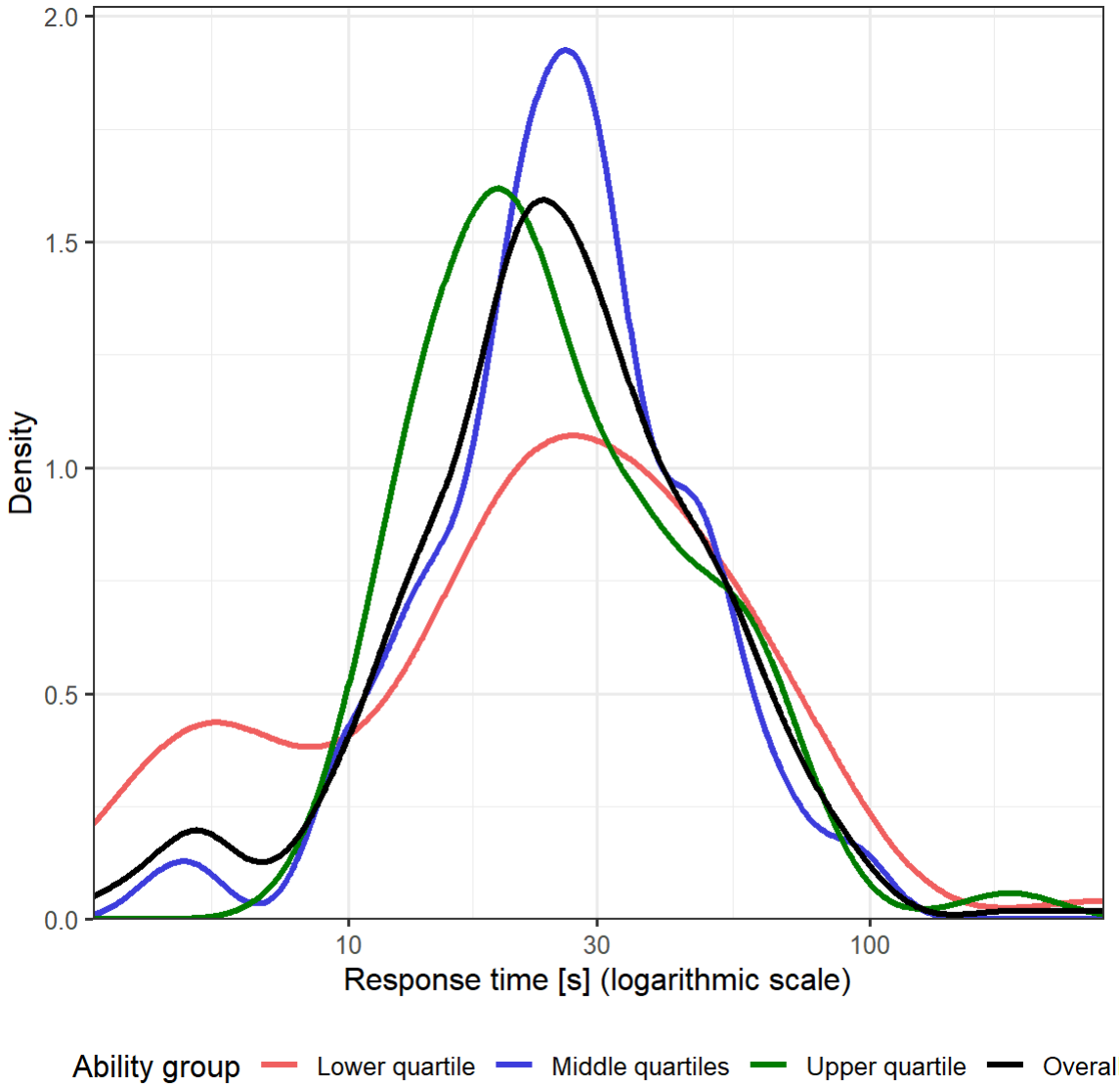



```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



jest cięższy od powietrza. Hel nagrzewa się szybciej niż powietrze. Hel przewodzi prąd l

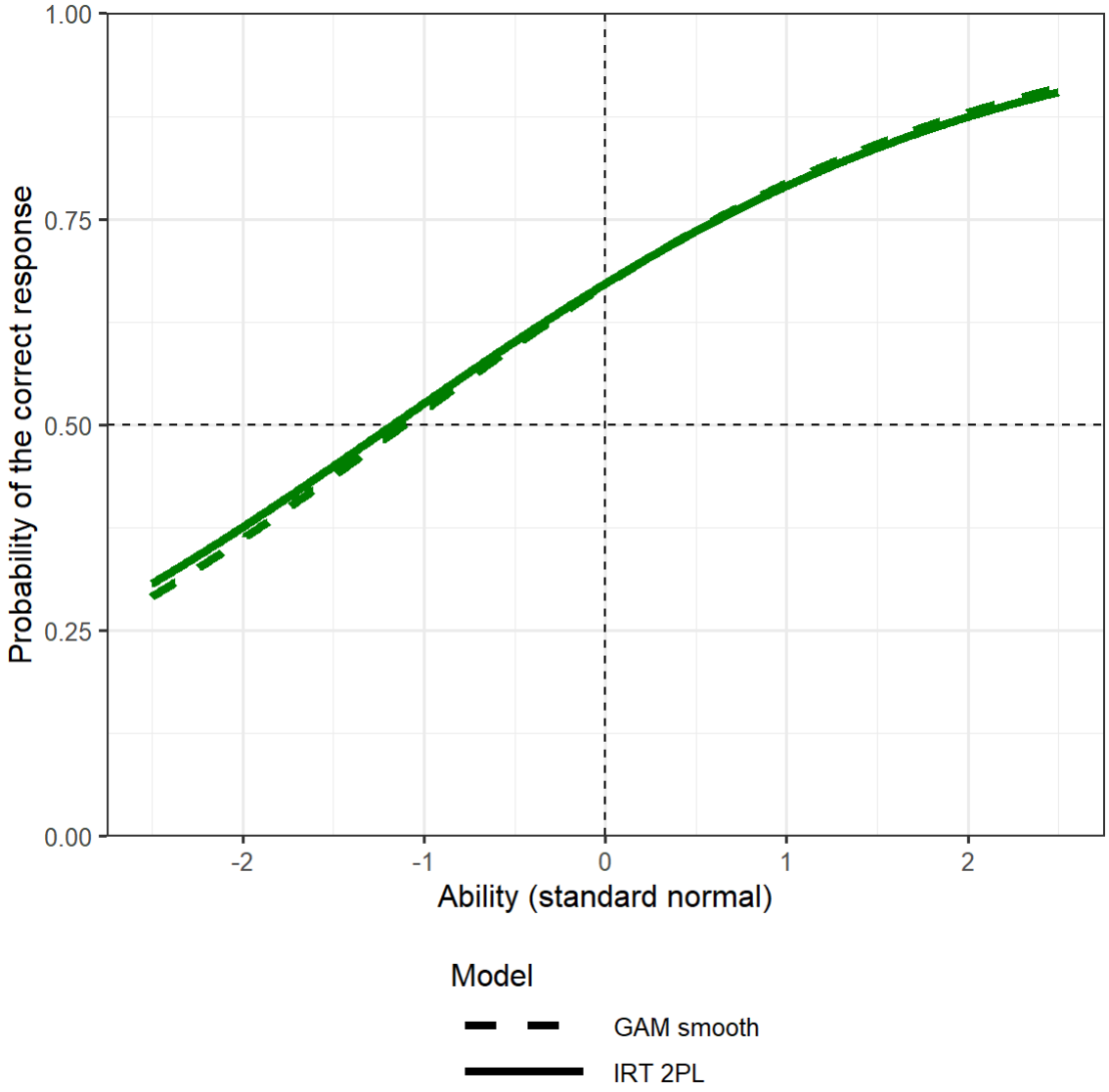
10.2.38.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.2	13.3	25.4	40.9	280.7	3.1	0.9
Middle quartiles	4.2	19.2	26.2	34.4	99.8	3.2	0.6
Upper quartile	9.5	16.5	21.8	37.3	186.1	3.2	0.6
Overall	3.2	16.9	24.9	37.6	280.7	3.2	0.7

10.2.39 S061105 (LLM)

10.2.39.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	65.46%	305	TIMSS	MC	0.27	0.24	0.61	−1.17	0.008

10.2.39.2 Gender DIF:

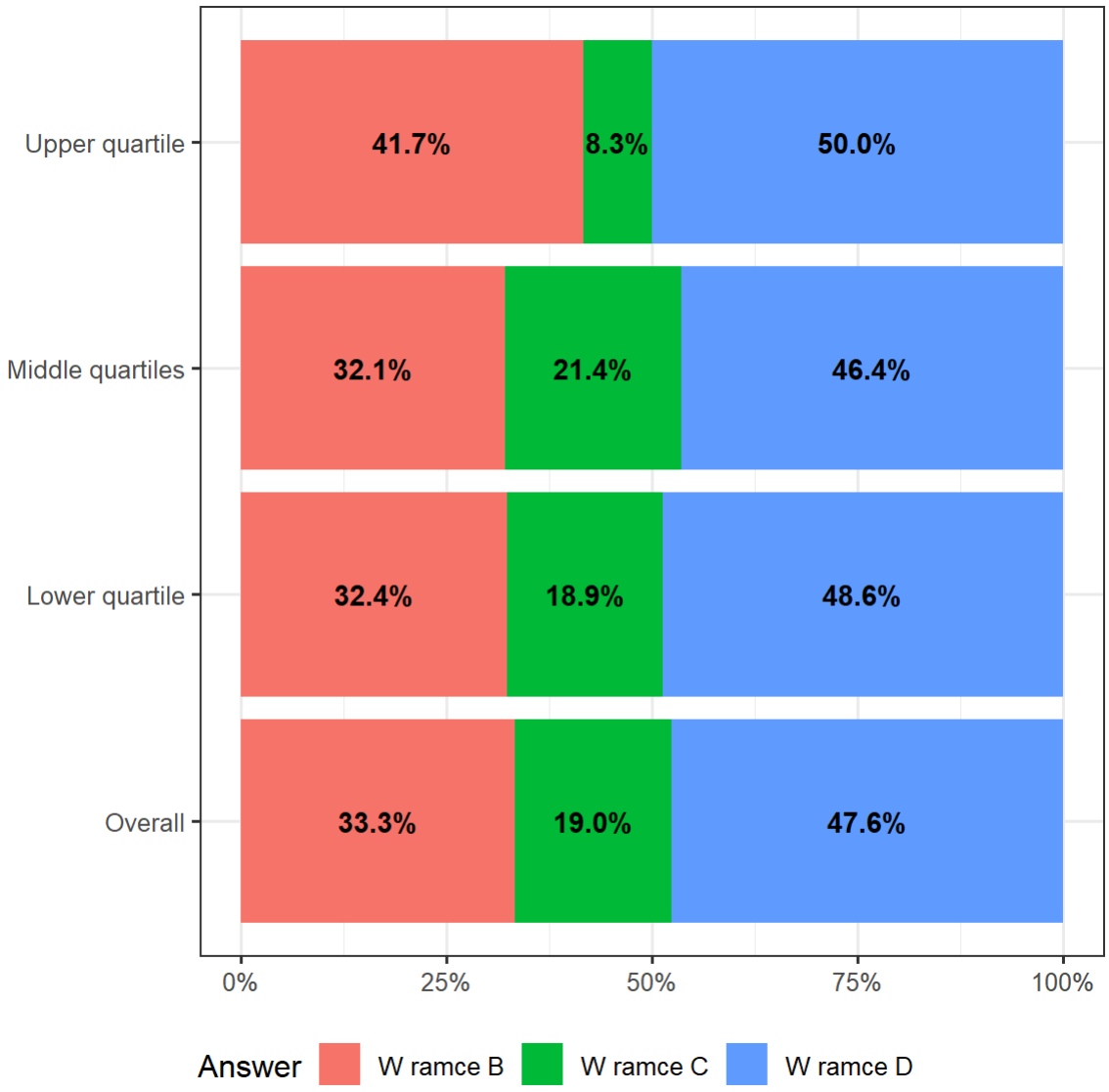
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
8.85	0.012	0.996	0.026	N/A

10.2.39.3 Frequency of answers:

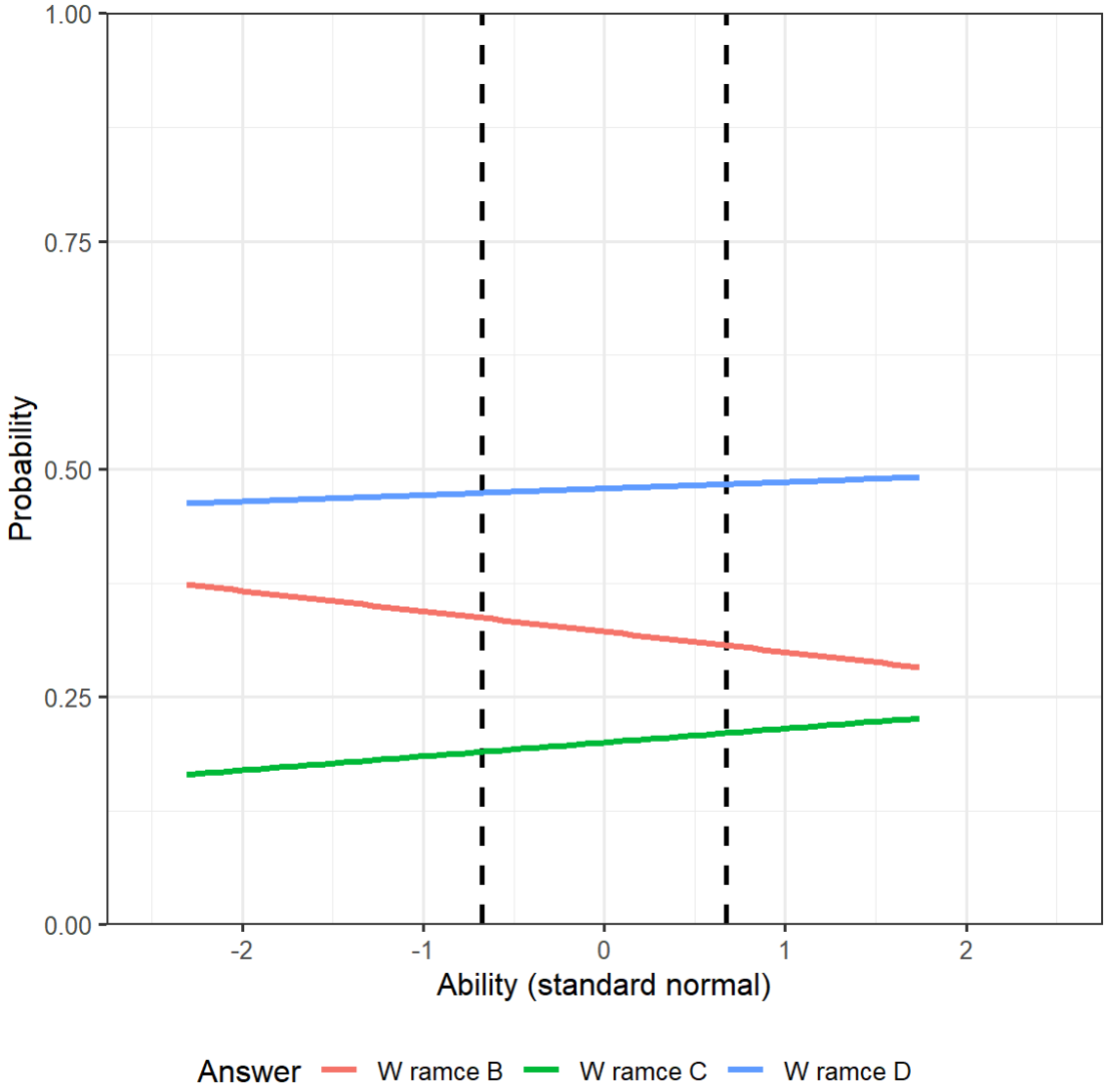
Answer	Frequency	Percent
W ramce A (correct)	199	65.5%
W ramce B	35	11.5%
W ramce C	20	6.6%
W ramce D	50	16.4%
Overall	304	100.0%

10.2.39.4 Distractors functioning:

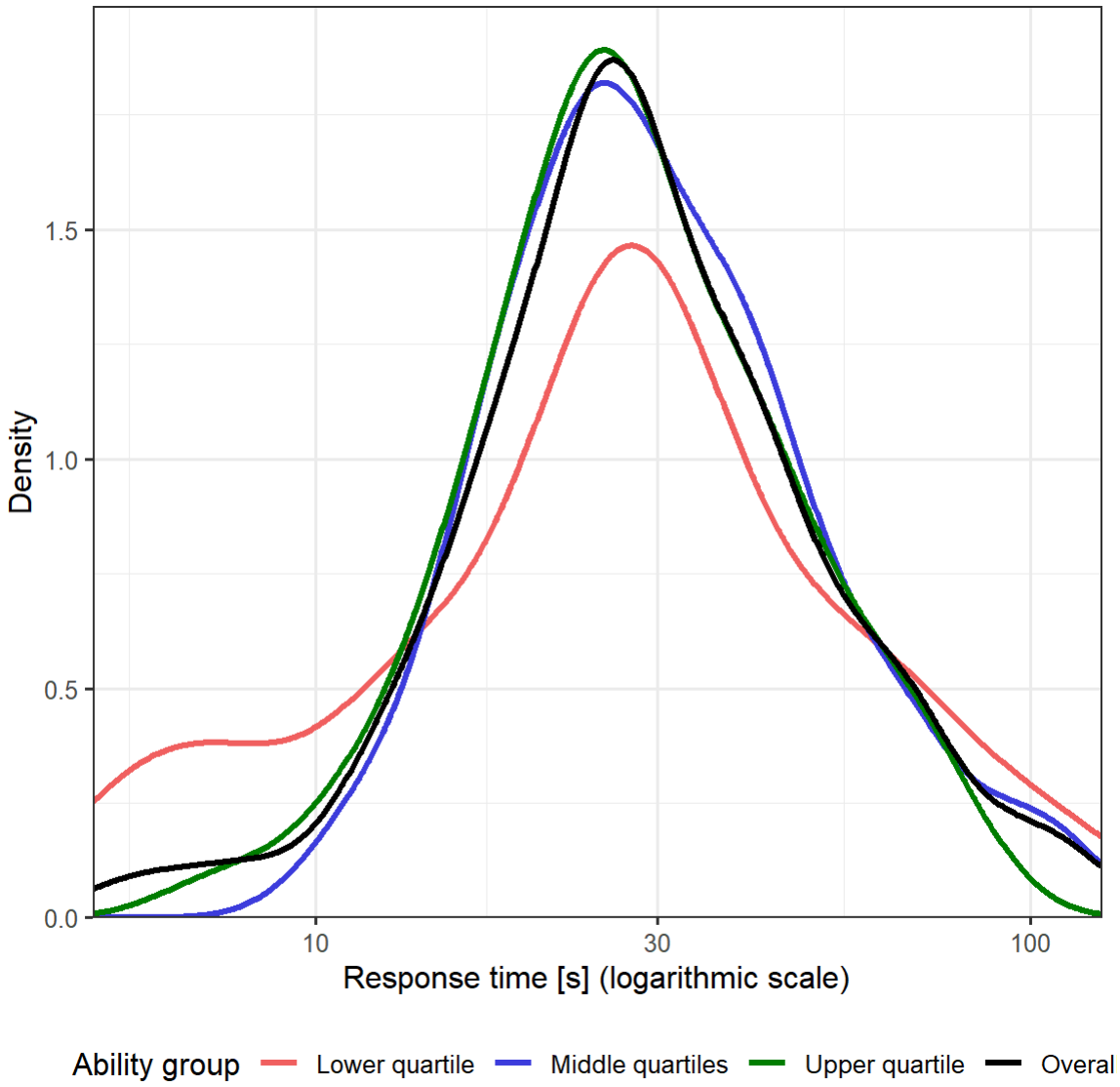
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
W ramce B	12	18	5	35	32.4%	32.1%	41.7%	33.3%
W ramce C	7	12	1	20	18.9%	21.4%	8.3%	19.0%
W ramce D	18	26	6	50	48.6%	46.4%	50.0%	47.6%
Overall	37	56	12	105	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



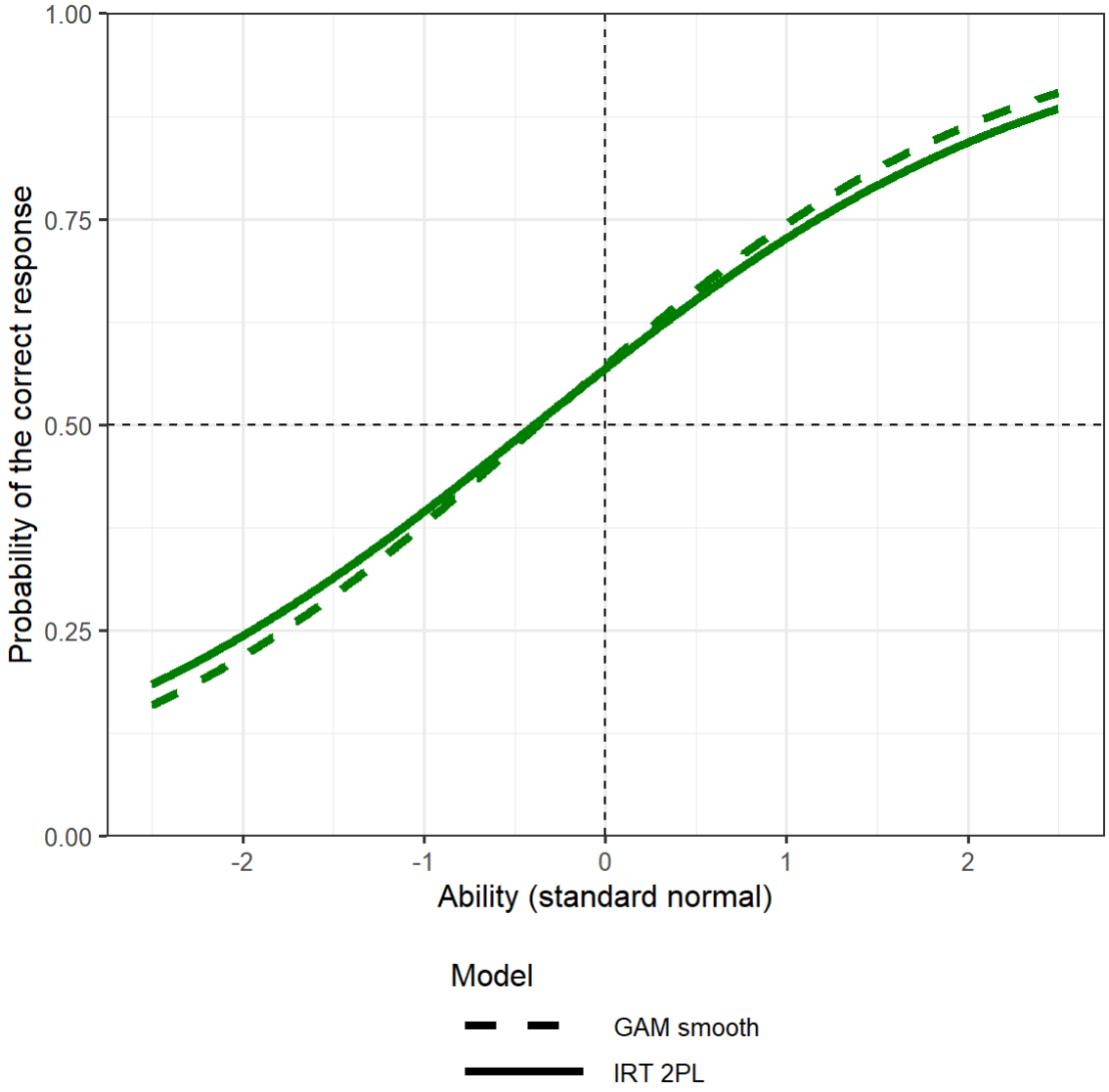
10.2.39.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.9	16.1	26.8	39.2	125.6	3.2	0.8
Middle quartiles	9.5	21.5	28.5	41.0	116.7	3.4	0.5
Upper quartile	7.3	20.2	27.1	38.5	81.4	3.3	0.5
Overall	4.9	20.1	27.3	40.4	125.6	3.3	0.6

10.2.40 S051114 (TIMSS)

10.2.40.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	56.71%	299	TIMSS	MC	0.33	0.37	0.70	−0.39	0.000

10.2.40.2 Gender DIF:

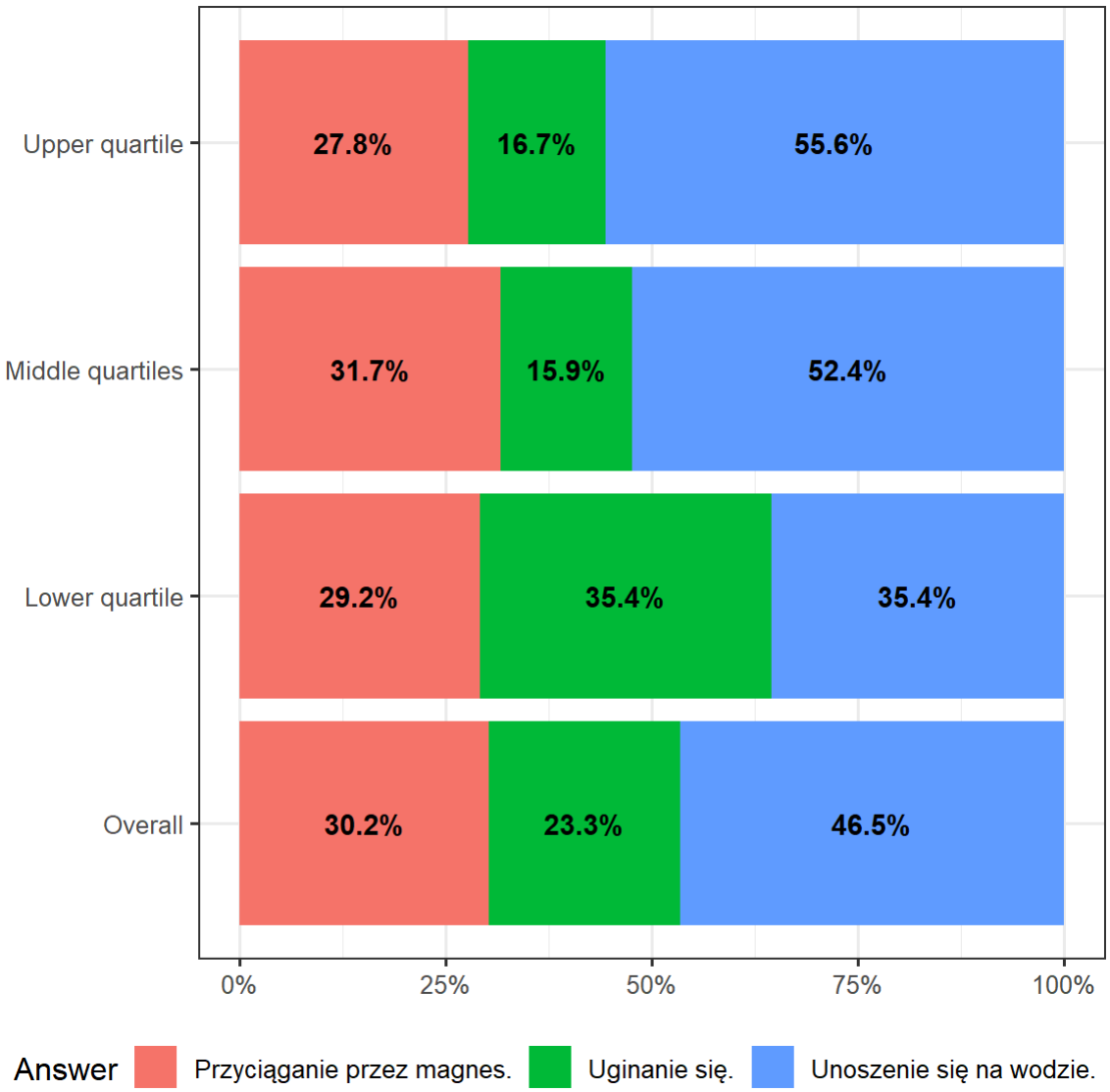
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.40	0.820	1.000	0.001	N/A

10.2.40.3 Frequency of answers:

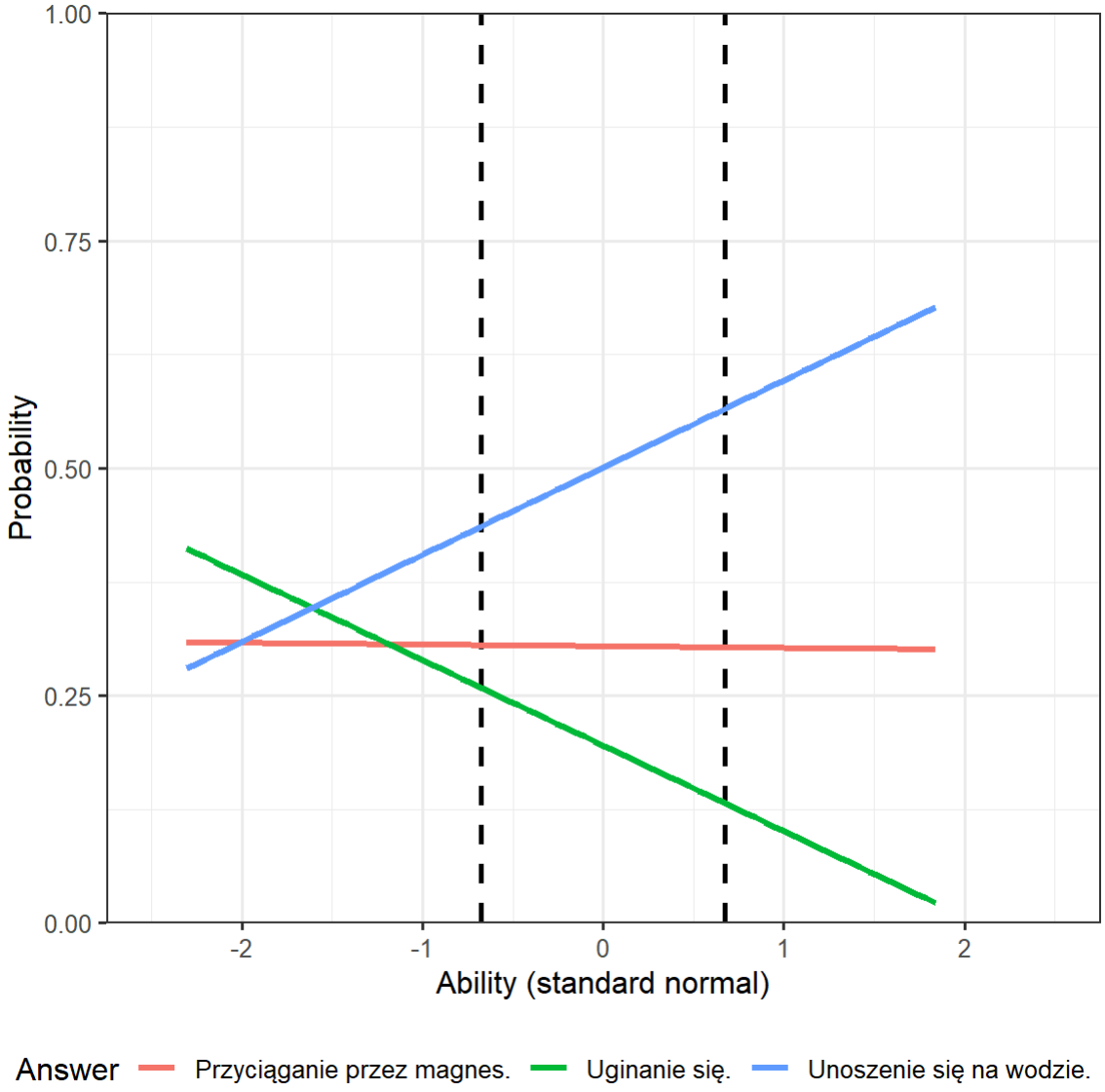
Answer	Frequency	Percent
Przepuszczanie światła. (correct)	169	56.7%
Przyciąganie przez magnes.	39	13.1%
Uginanie się.	30	10.1%
Unoszenie się na wodzie.	60	20.1%
Overall	298	100.0%

10.2.40.4 Distractors functioning:

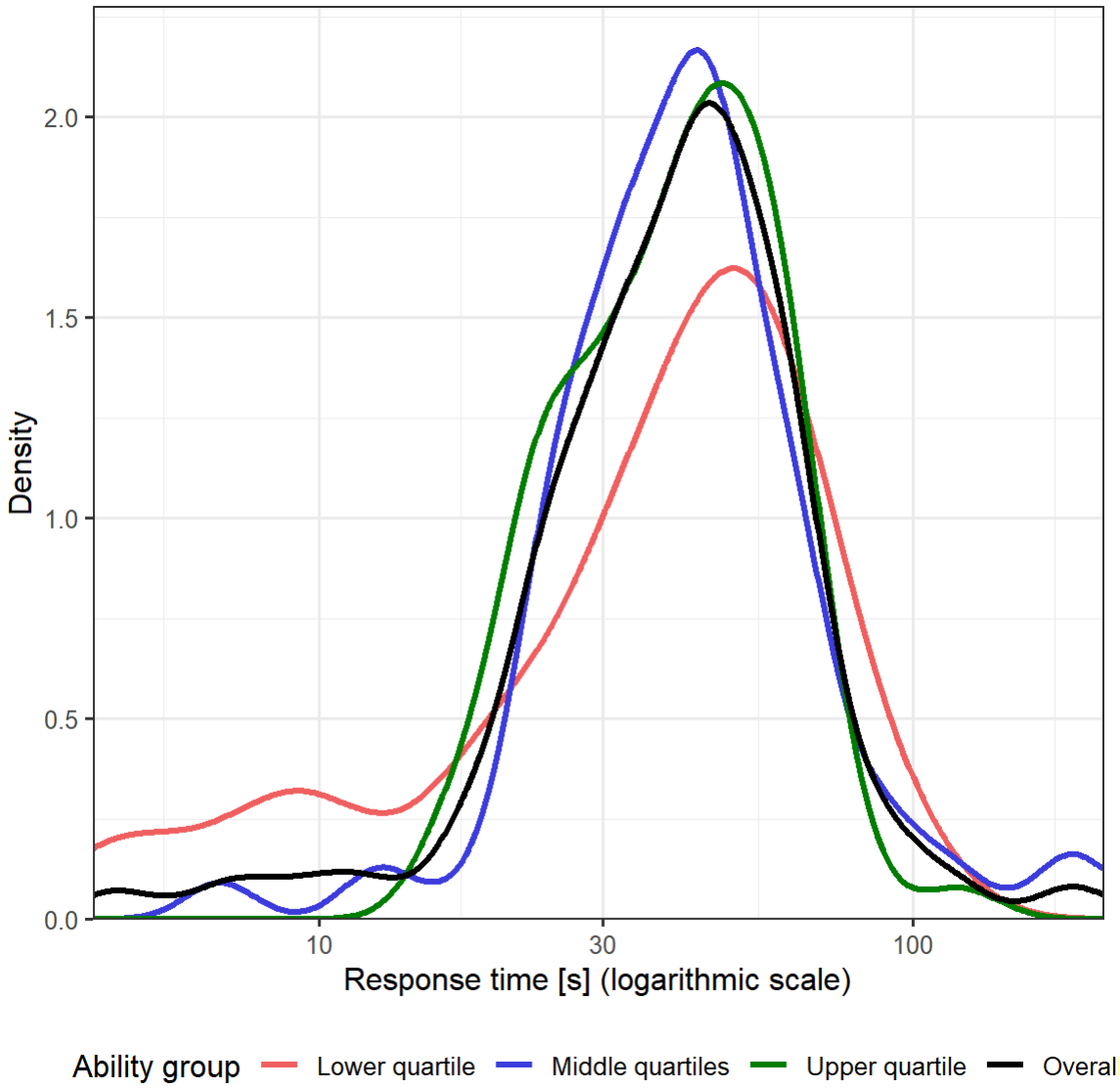
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Przyciąganie przez magnes.	14	20	5	39	29.2%	31.7%	27.8%	30.2%
Uginanie się.	17	10	3	30	35.4%	15.9%	16.7%	23.3%
Unoszenie się na wodzie.	17	33	10	60	35.4%	52.4%	55.6%	46.5%
Overall	48	63	18	129	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



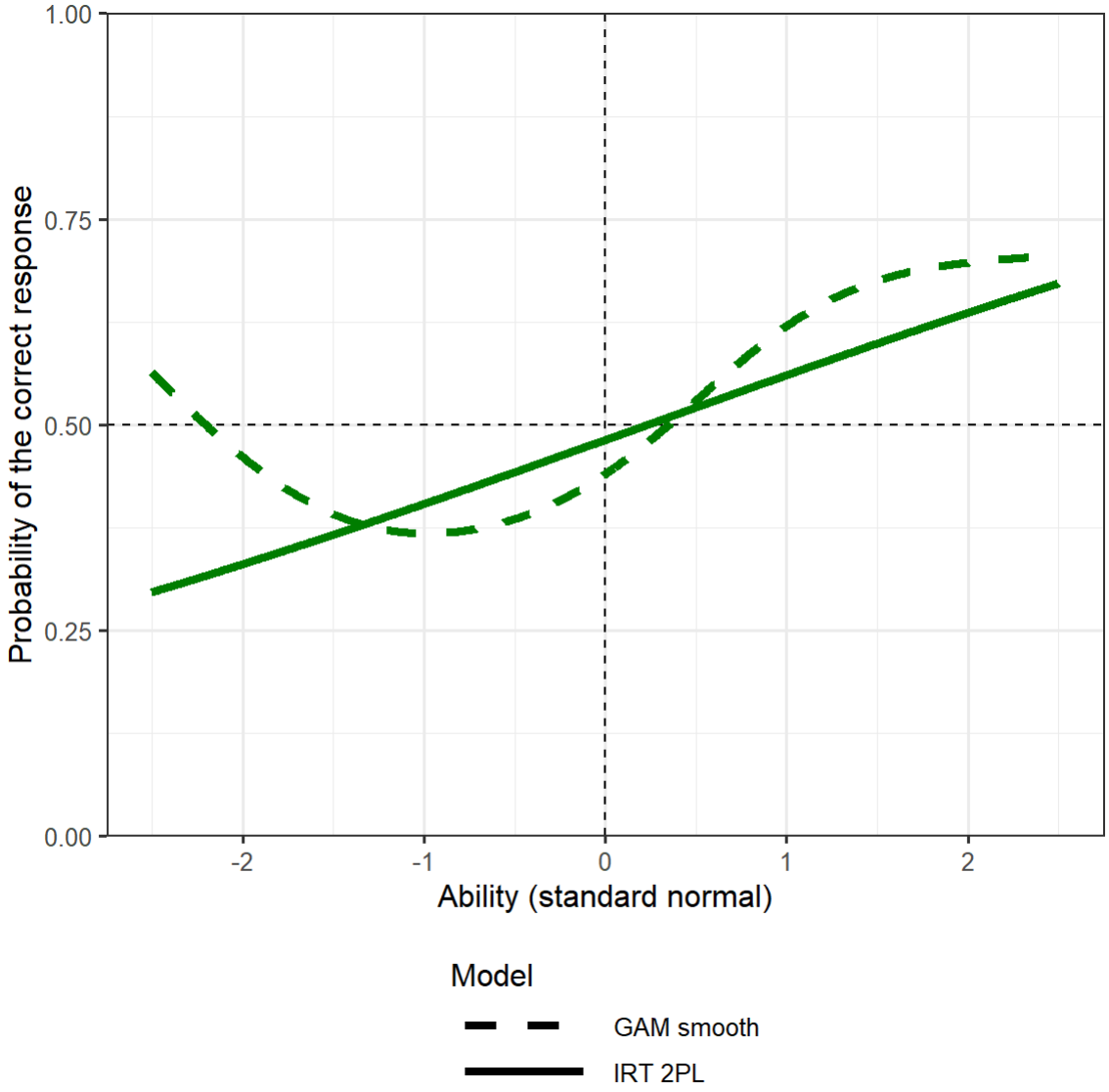
10.2.40.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.2	24.1	40.9	56.4	99.9	3.5	0.8
Middle quartiles	6.7	30.8	41.7	53.3	208.7	3.7	0.5
Upper quartile	16.2	29.4	41.4	53.5	120.3	3.7	0.4
Overall	4.2	29.4	41.5	54.4	208.7	3.6	0.6

10.2.41 S890016 (LLM)

10.2.41.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	48.28%	319	LLM complex setup	MC	0.16	0.06	0.32	0.23	0.02

10.2.41.2 Gender DIF:

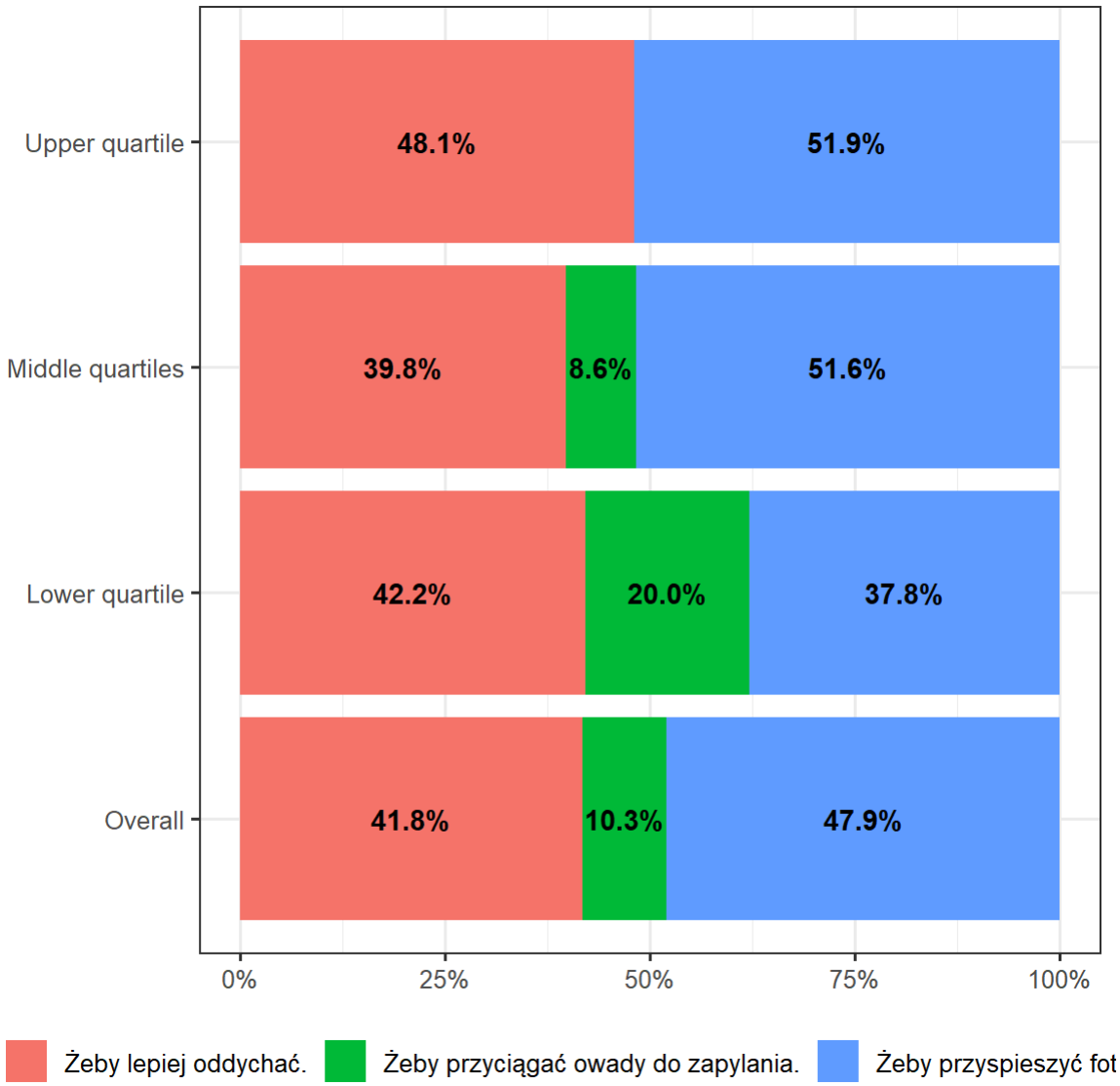
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.61	0.164	1.000	0.010	N/A

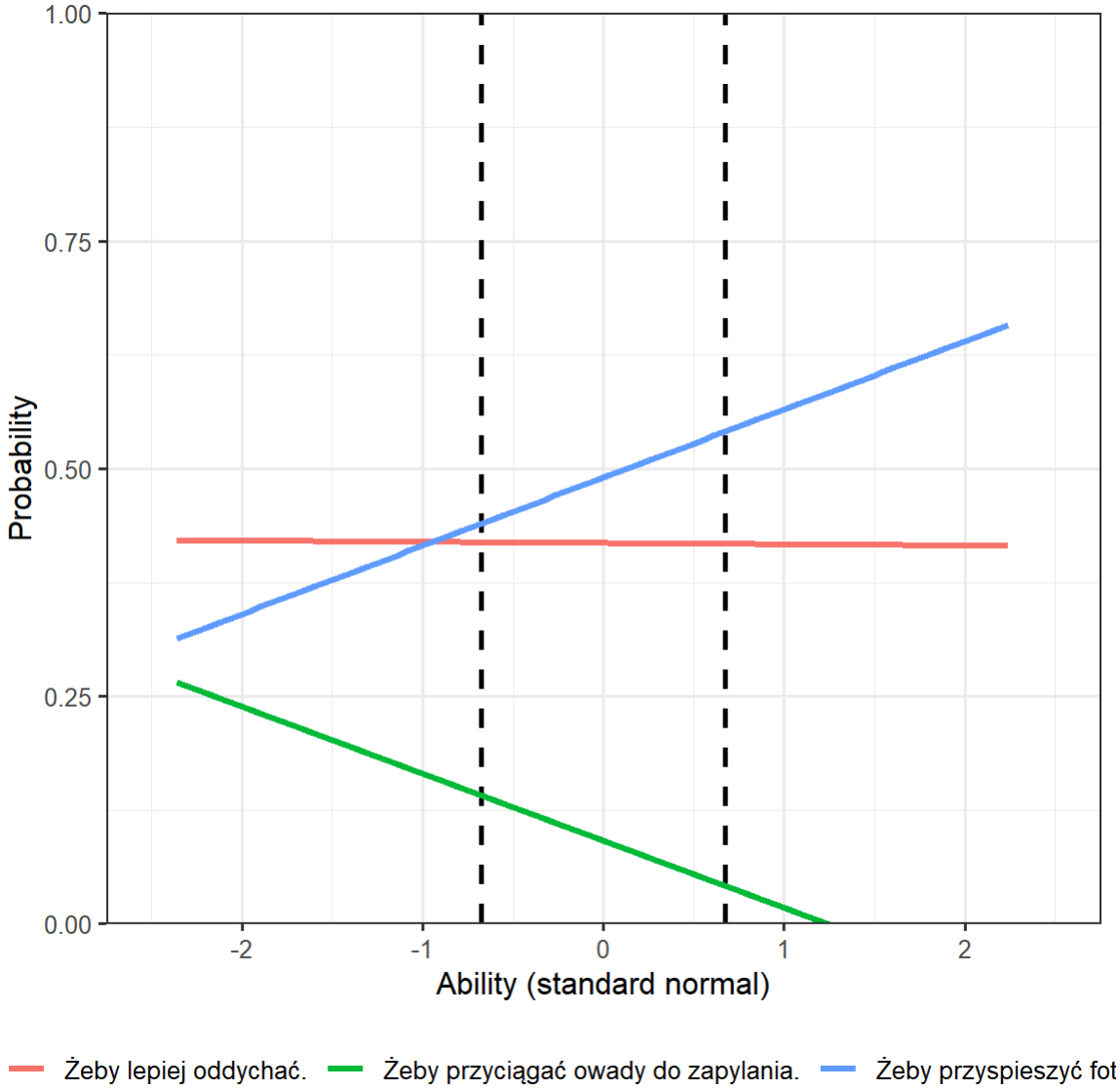
10.2.41.3 Frequency of answers:

Answer	Frequency	Percent
Żeby lepiej oddychać.	69	21.6%
Żeby magazynować więcej wody. (correct)	154	48.3%
Żeby przyciągać owady do zapylania.	17	5.3%
Żeby przyspieszyć fotosyntezę.	79	24.8%
Overall	319	100.0%

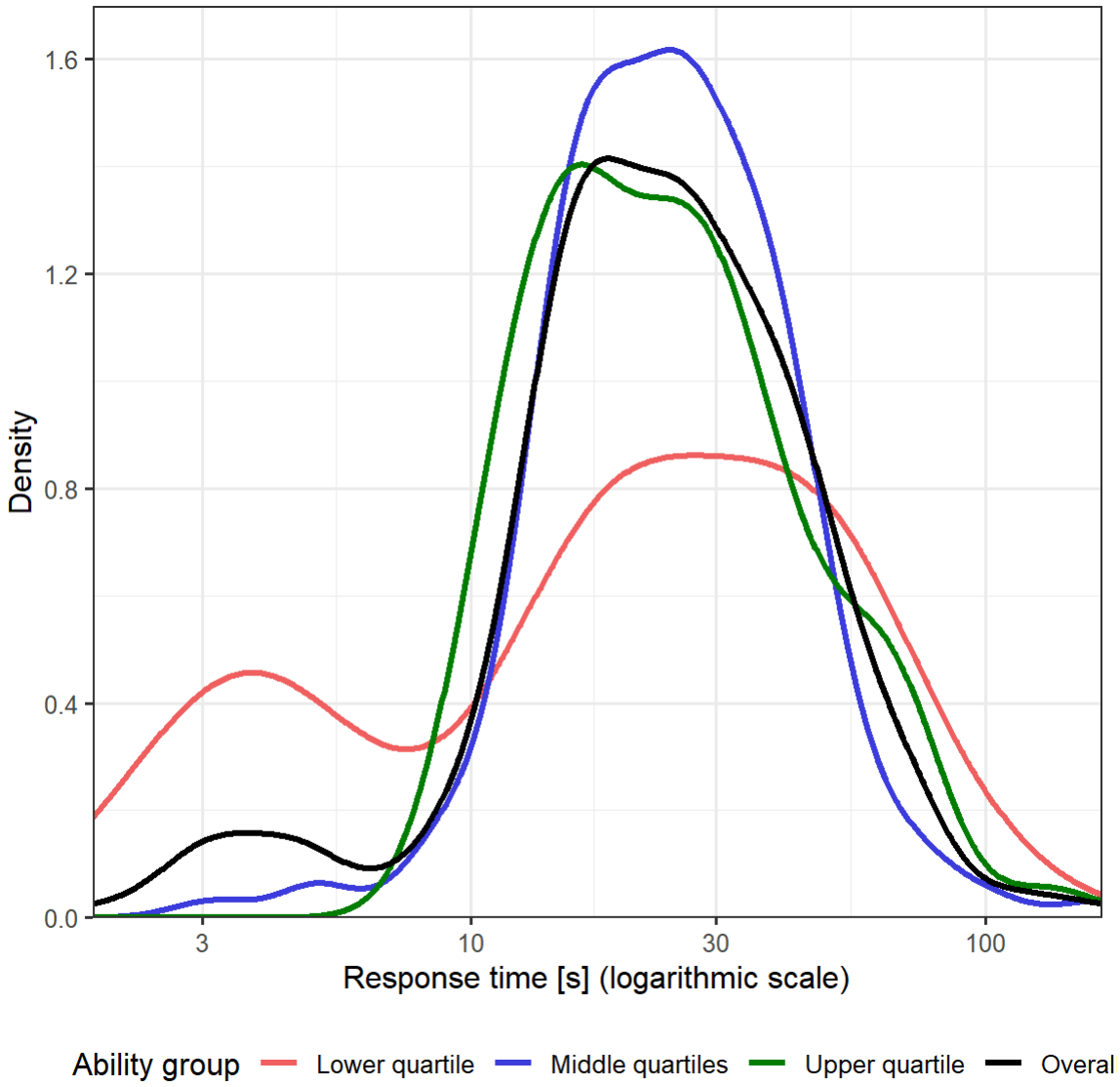
10.2.41.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Żeby lepiej oddychać.	19	37	13	69	42.2%	39.8%	48.1%	41.8%
Żeby przyciągać owady do zapylania.	9	8	0	17	20.0%	8.6%	0.0%	10.3%
Żeby przyspieszyć fotosyntezę.	17	48	14	79	37.8%	51.6%	51.9%	47.9%
Overall	45	93	27	165	100.0%	100.0%	100.0%	100.0%


Upper quartile
Middle quartiles
Lower quartile
Overall
48.1%
51.9%
39.8%
8.6%
51.6%
42.2%
20.0%
37.8%
41.8%
10.3%
47.9%
0%
25%
50%
75%
100%
Żeby lepiej oddychać.
Żeby przyciągać owady do zapylania.
Żeby przyspieszyć fot



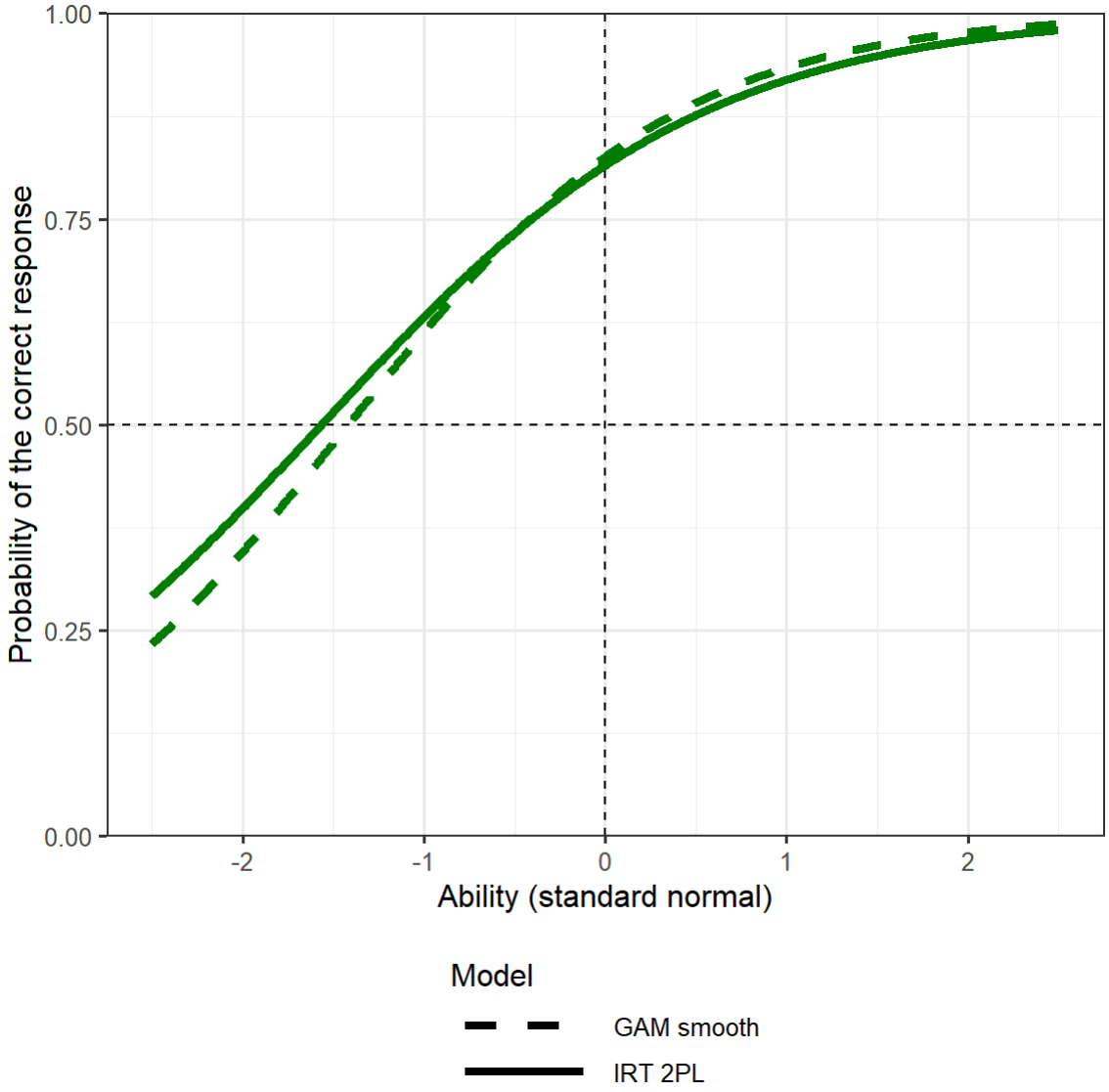
10.2.41.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	1.8	7.1	20.3	44.2	116.8	2.9	1.1
Middle quartiles	3.1	16.8	24.5	34.9	168.2	3.2	0.5
Upper quartile	9.1	15.4	23.8	35.6	134.3	3.2	0.6
Overall	1.8	15.7	23.6	36.1	168.2	3.1	0.7

10.2.42 SP71402 (TIMSS)

10.2.42.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	78.52%	299	TIMSS	MC	0.37	0.36	0.95	−1.56	0.005

10.2.42.2 Gender DIF:

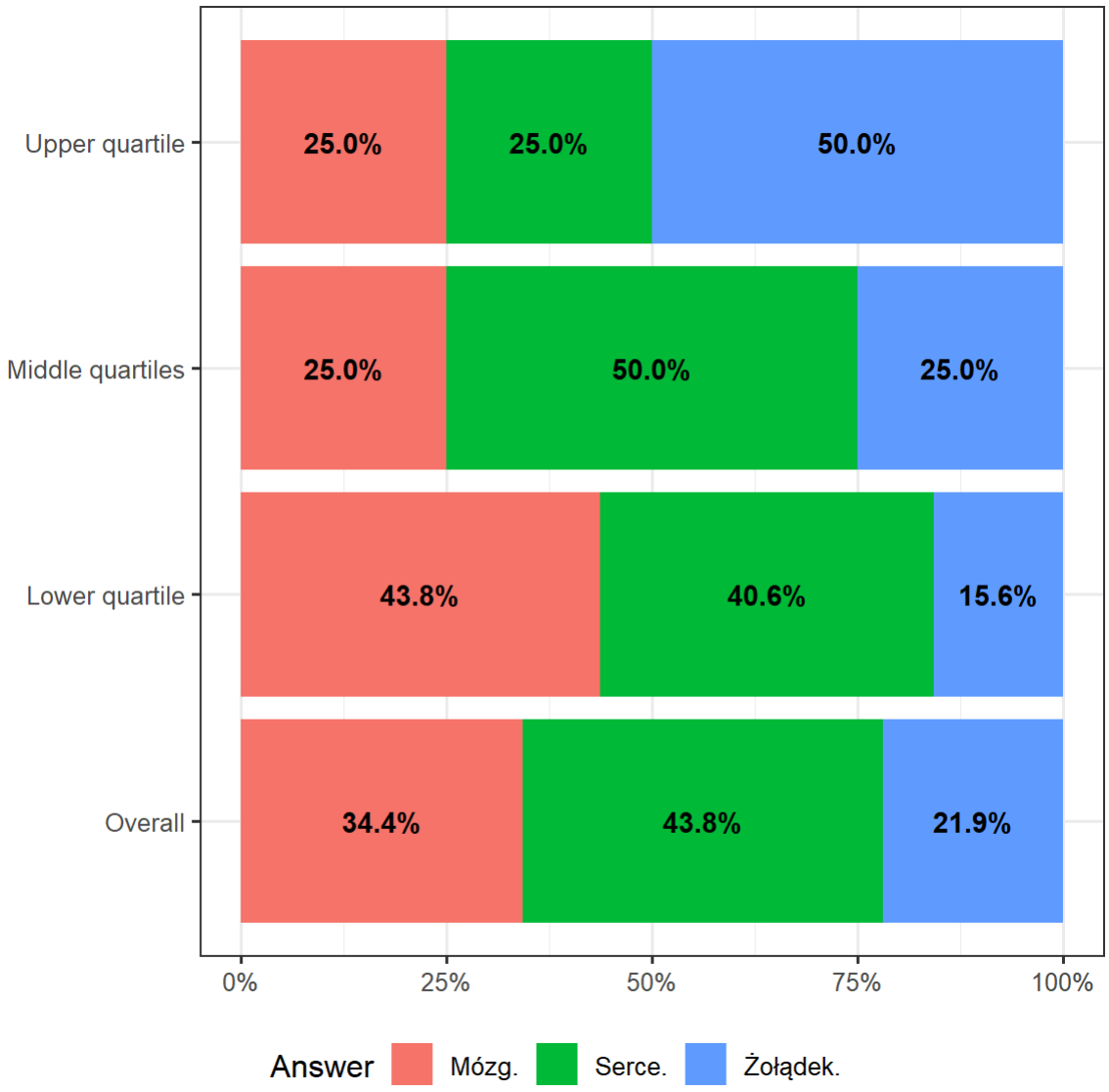
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.37	0.504	1.000	0.005	N/A

10.2.42.3 Frequency of answers:

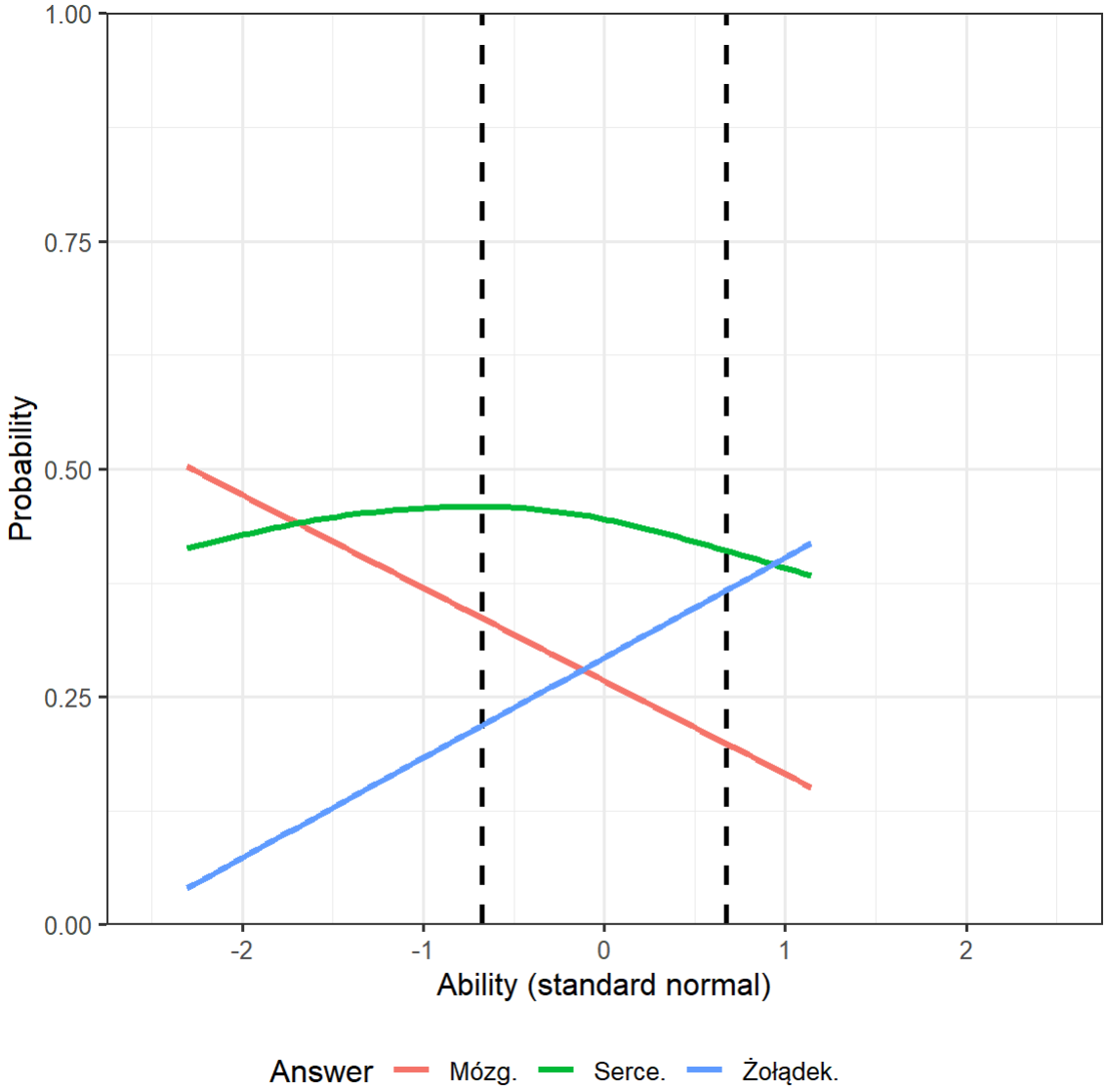
Answer	Frequency	Percent
Mózg.	22	7.4%
Płuca. (correct)	234	78.5%
Serce.	28	9.4%
Żołądek.	14	4.7%
Overall	298	100.0%

10.2.42.4 Distractors functioning:

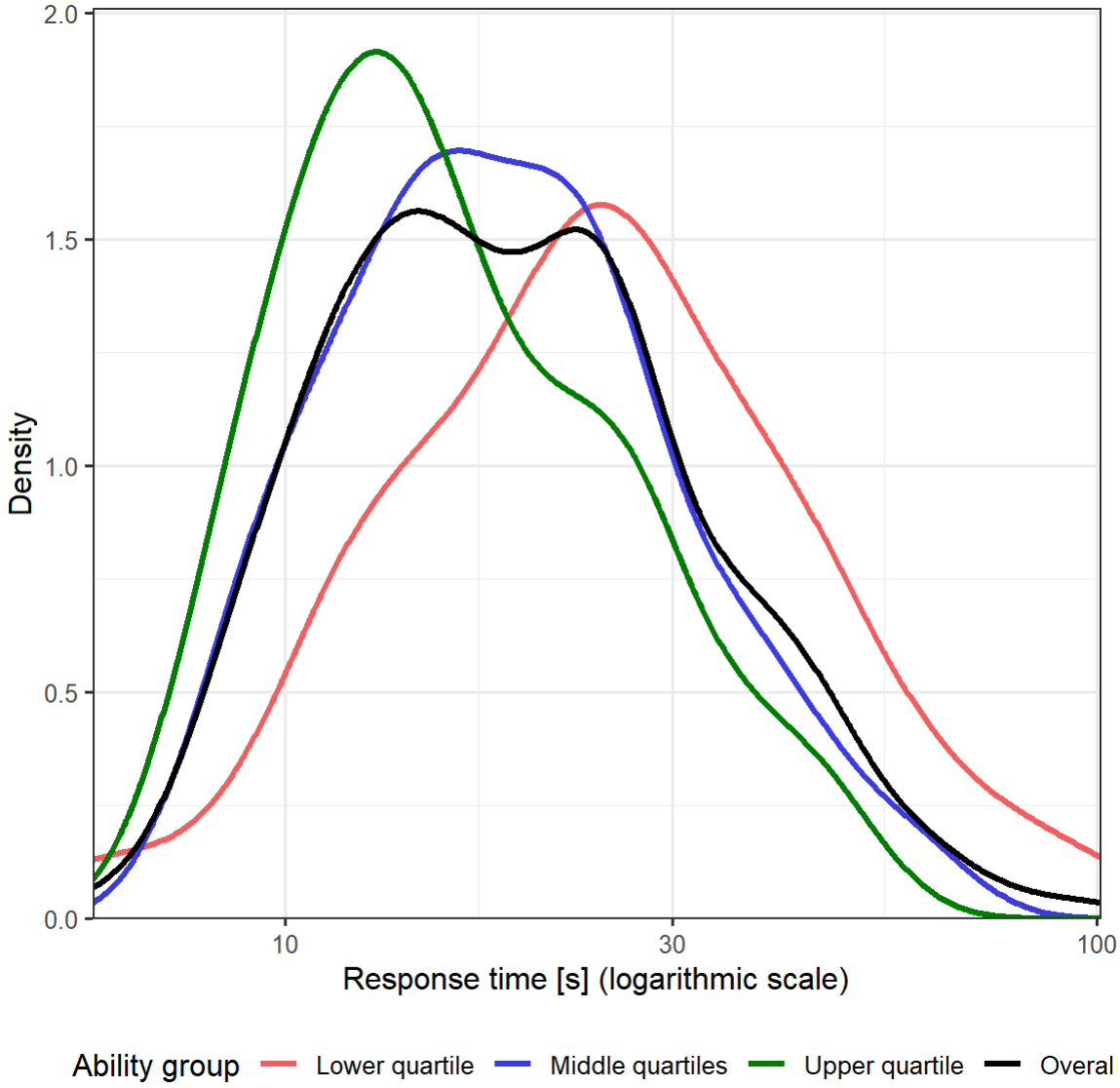
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Mózg.	14	7	1	22	43.8%	25.0%	25.0%	34.4%
Serce.	13	14	1	28	40.6%	50.0%	25.0%	43.8%
Żołądek.	5	7	2	14	15.6%	25.0%	50.0%	21.9%
Overall	32	28	4	64	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



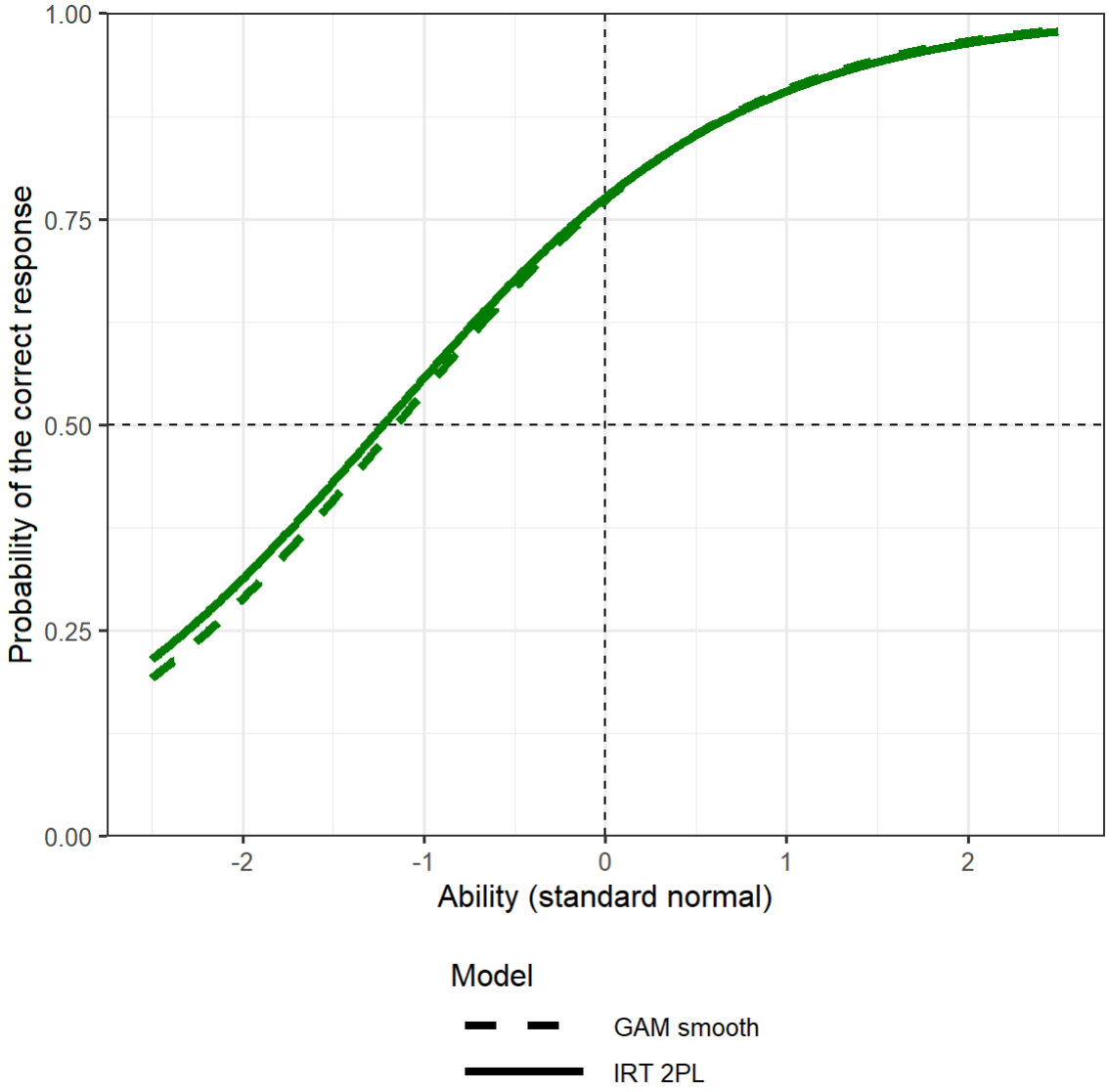
10.2.42.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.8	16.5	24.6	36.1	101.1	3.2	0.6
Middle quartiles	7.5	13.1	17.7	25.6	64.4	2.9	0.5
Upper quartile	7.5	11.4	14.7	23.5	48.9	2.8	0.5
Overall	5.8	12.9	18.4	26.6	101.1	3.0	0.5

10.2.43 S100029 (LLM)

10.2.43.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS a generated items Discrimination	Item difficulty	RMSEA
1	73.36%	305	LLM simple setup	MC	0.38	0.28	1.01	−1.22	0.018

10.2.43.2 Gender DIF:

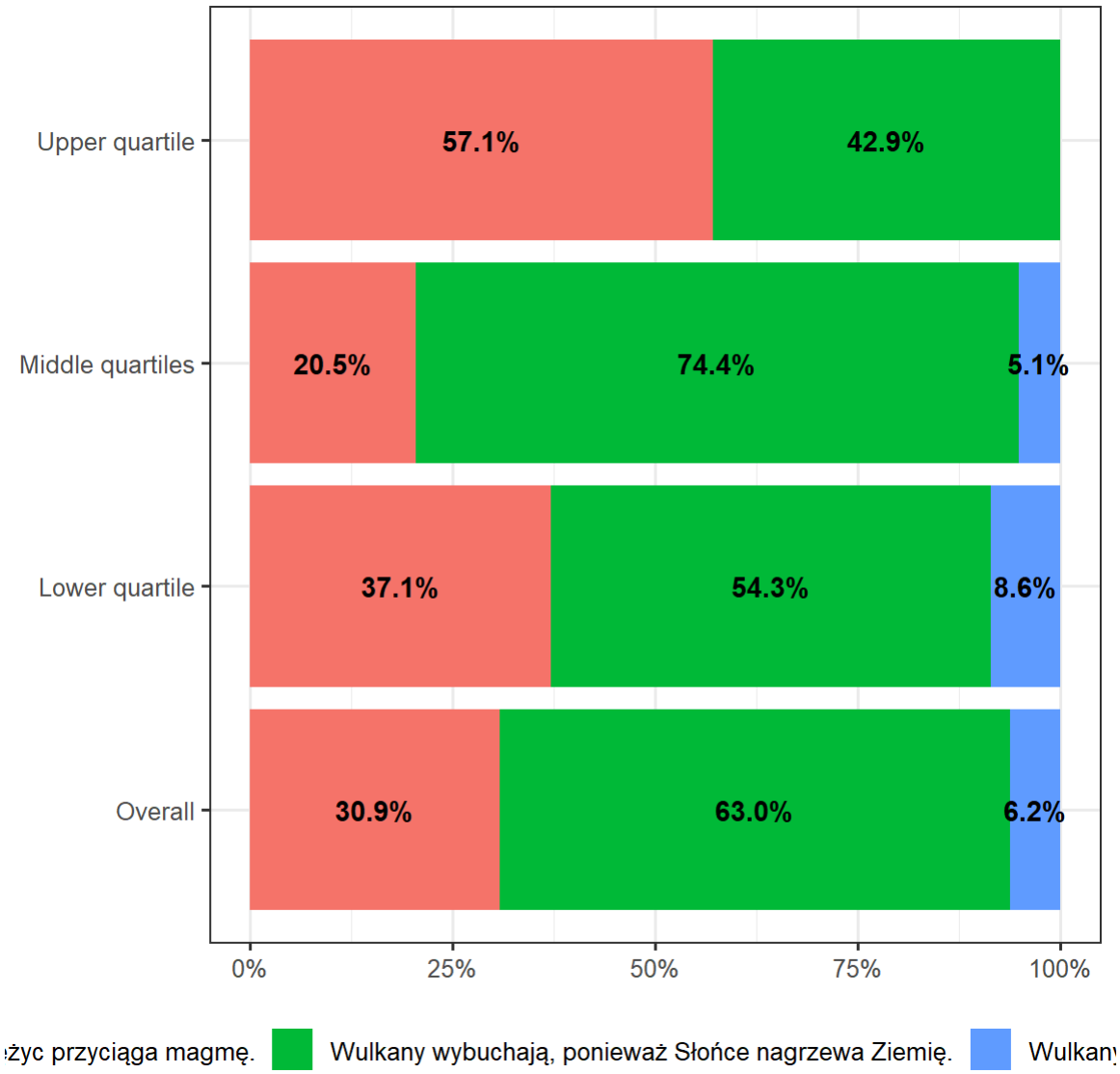
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.34	0.512	1.000	0.004	N/A

10.2.43.3 Frequency of answers:

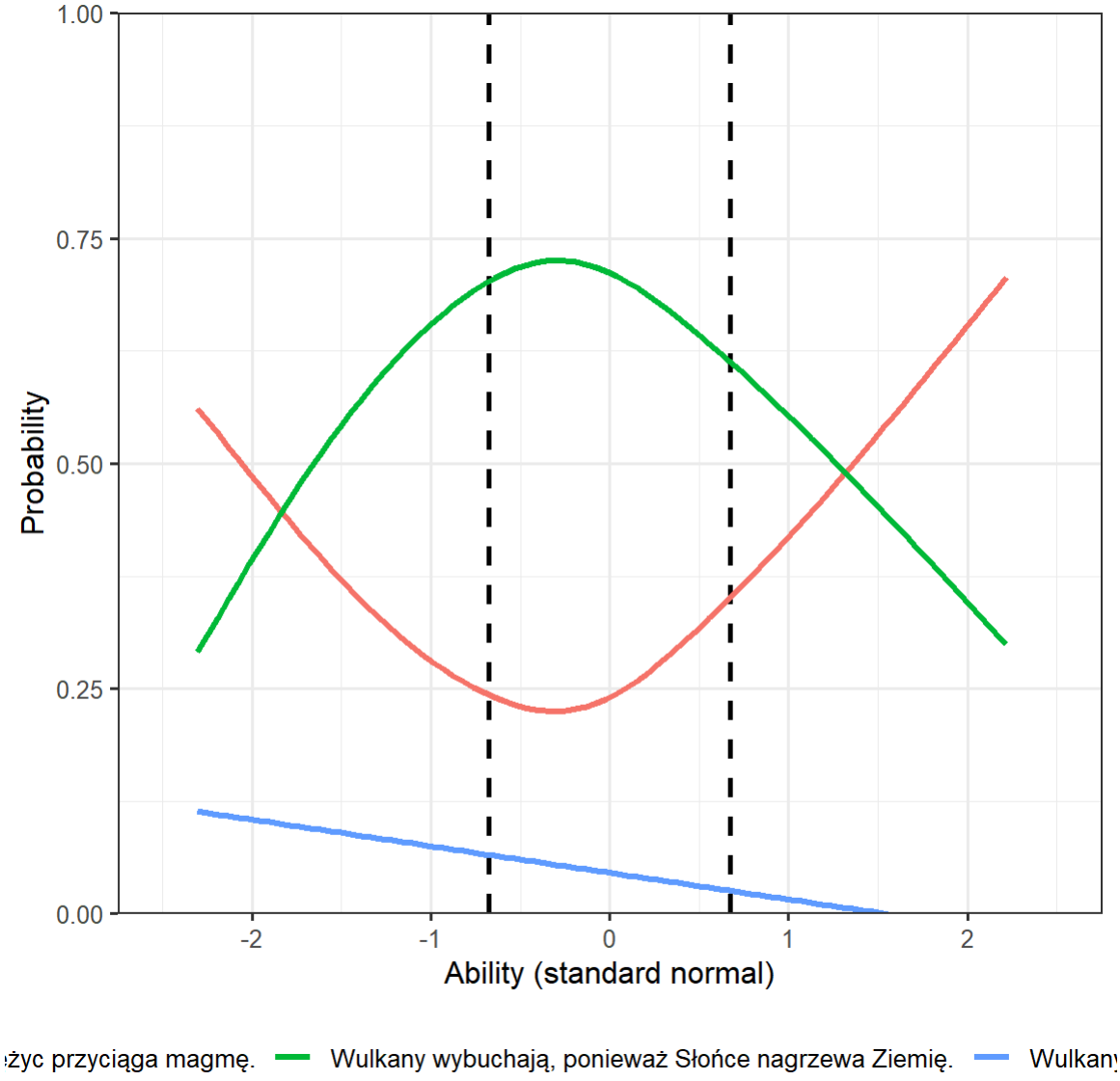
Answer	Frequency	Percent
Wulkany wybuchają, ponieważ Słońce nagrzewa Ziemię.	51	16.8%
Wulkany wybuchają, ponieważ gorąca magma podnosi się na powierzchnię Ziemi. (correct)	223	73.4%
Wulkany wybuchają, ponieważ księżyc przyciąga magmę.	25	8.2%
Wulkany wybuchają, ponieważ wiatr wieje mocno na Ziemi.	5	1.6%
Overall	304	100.0%

10.2.43.4 Distractors functioning:

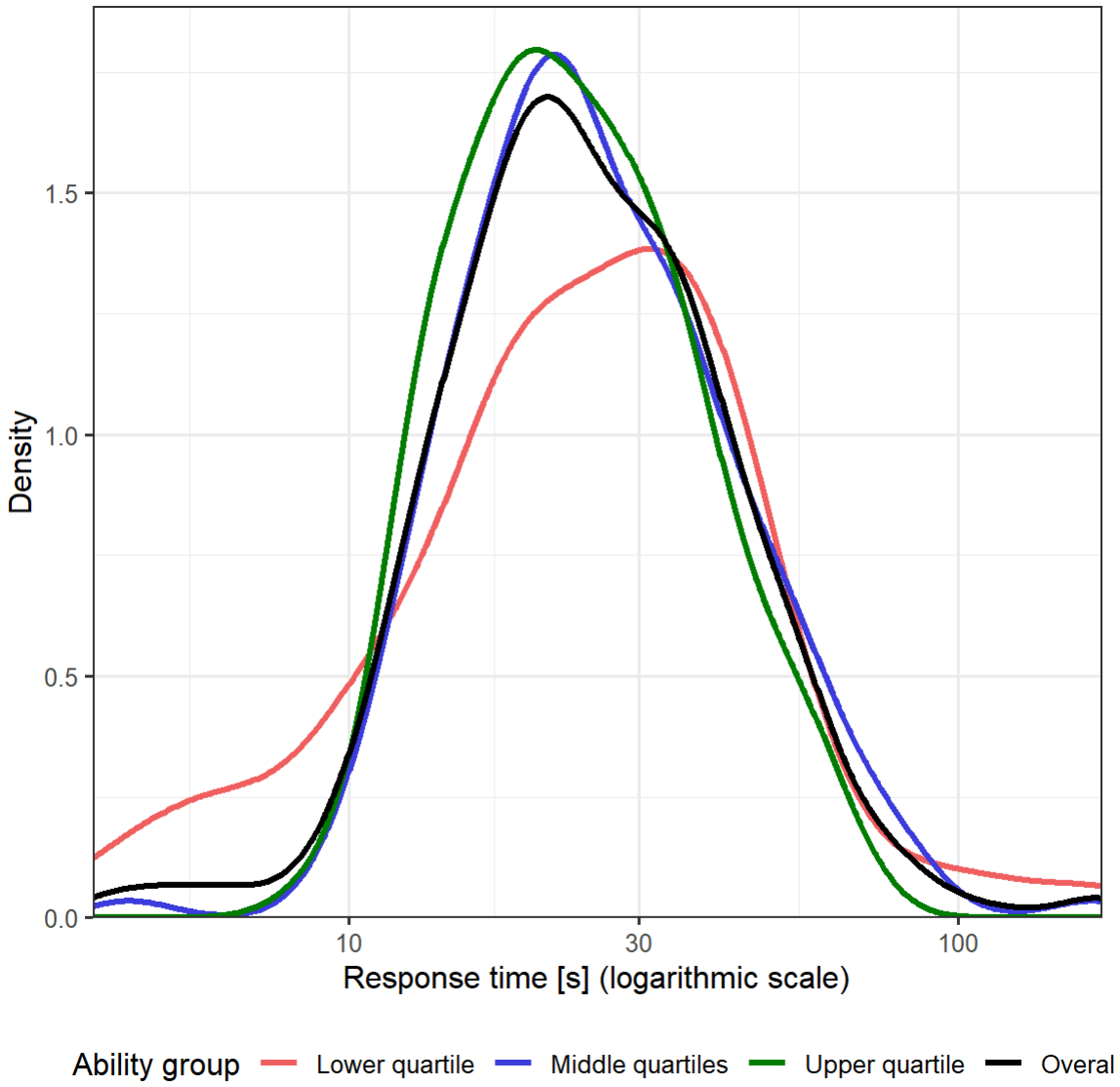
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Wulkany wybuchają, ponieważ Słońce nagrzewa Ziemię.	19	29	3	51	54.3%	74.4%	42.9%	63.0%
Wulkany wybuchają, ponieważ księżyc przyciąga magmę.	13	8	4	25	37.1%	20.5%	57.1%	30.9%
Wulkany wybuchają, ponieważ wiatr wieje mocno na Ziemi.	3	2	0	5	8.6%	5.1%	0.0%	6.2%
Overall	35	39	7	81	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



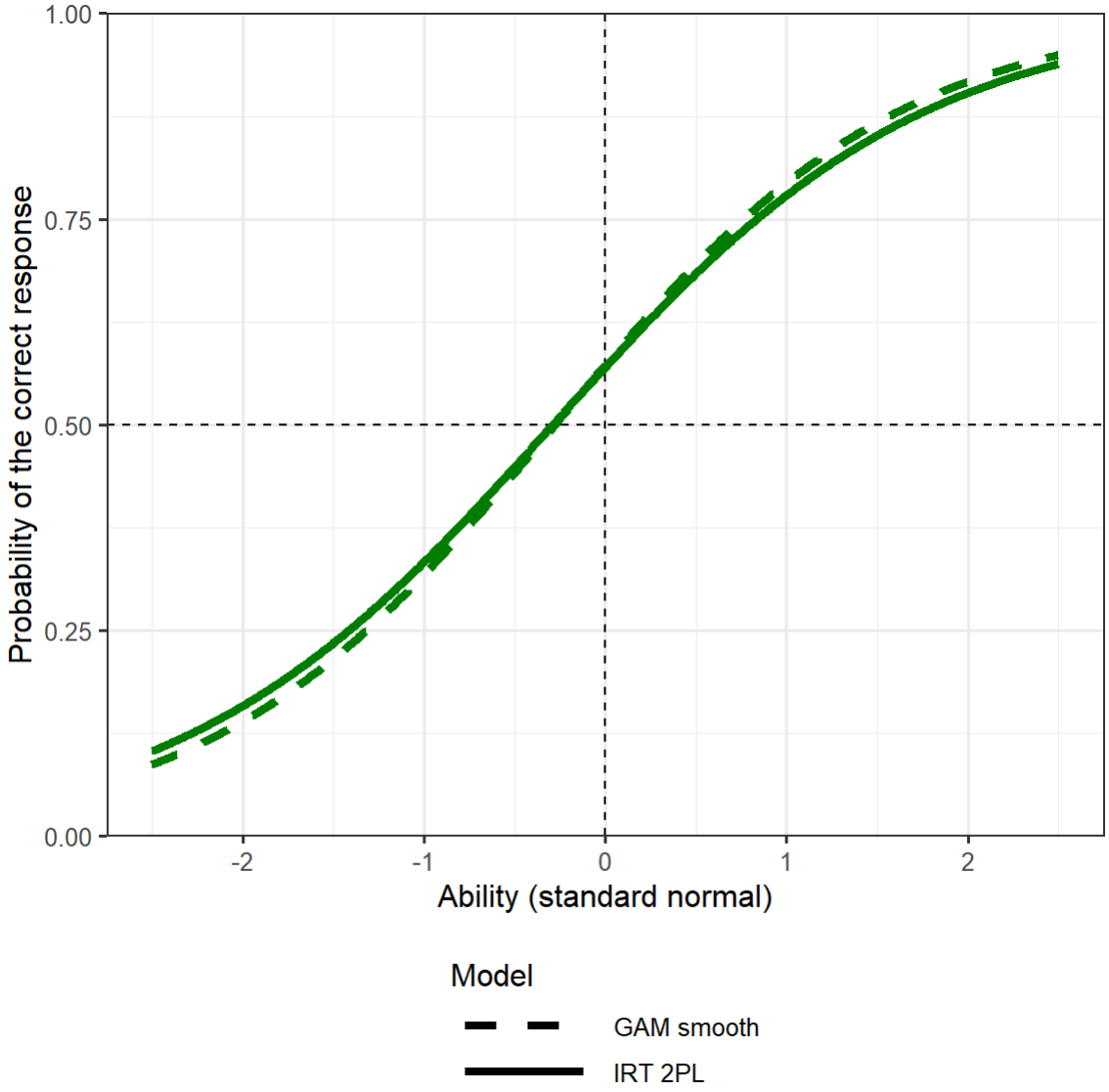
10.2.43.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.8	16.3	25.3	36.3	172.0	3.2	0.7
Middle quartiles	4.4	17.7	24.2	34.7	166.3	3.2	0.5
Upper quartile	9.3	16.9	22.7	32.4	64.7	3.2	0.4
Overall	3.8	17.2	24.1	34.5	172.0	3.2	0.6

10.2.44 S100020 (LLM)

10.2.44.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	56.11%	319	LLM simple setup	MC	0.41	0.27	0.98	−0.29	0.009

10.2.44.2 Gender DIF:

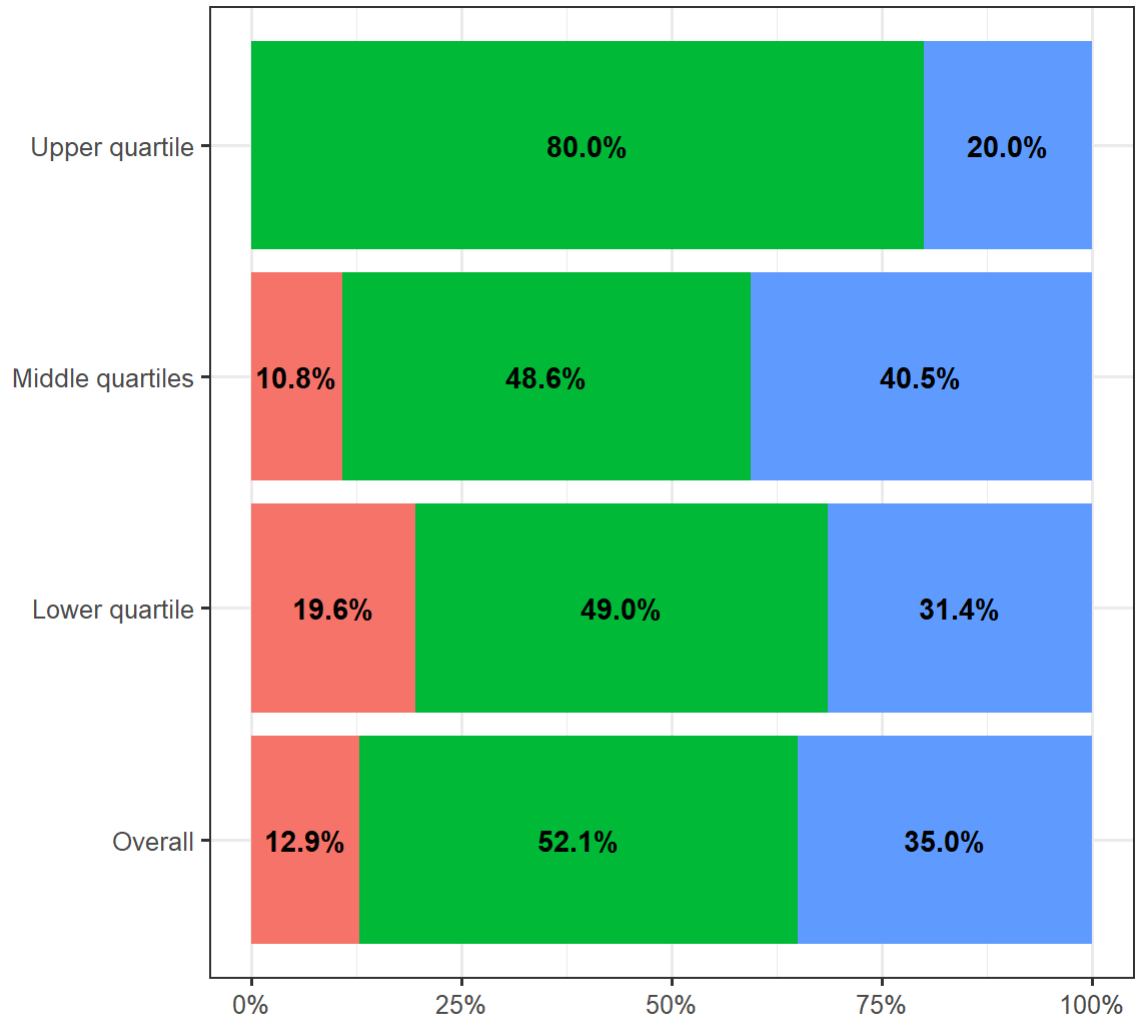
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.02	0.992	1.000	0.000	N/A

10.2.44.3 Frequency of answers:

Answer	Frequency	Percent
Drzewa te będą miały trudności z przetrwaniem i mogą zacząć wymierać. (correct)	179	56.1%
Drzewa te będą przenosić się do chłodniejszych obszarów, gdzie będą mogły przetrwać.	18	5.6%
Drzewa te będą przystosowywać się do nowych warunków, zmieniając swoje cechy genetyczne.	73	22.9%
Drzewa te będą rosły szybciej i stawały się silniejsze.	49	15.4%
Overall	319	100.0%

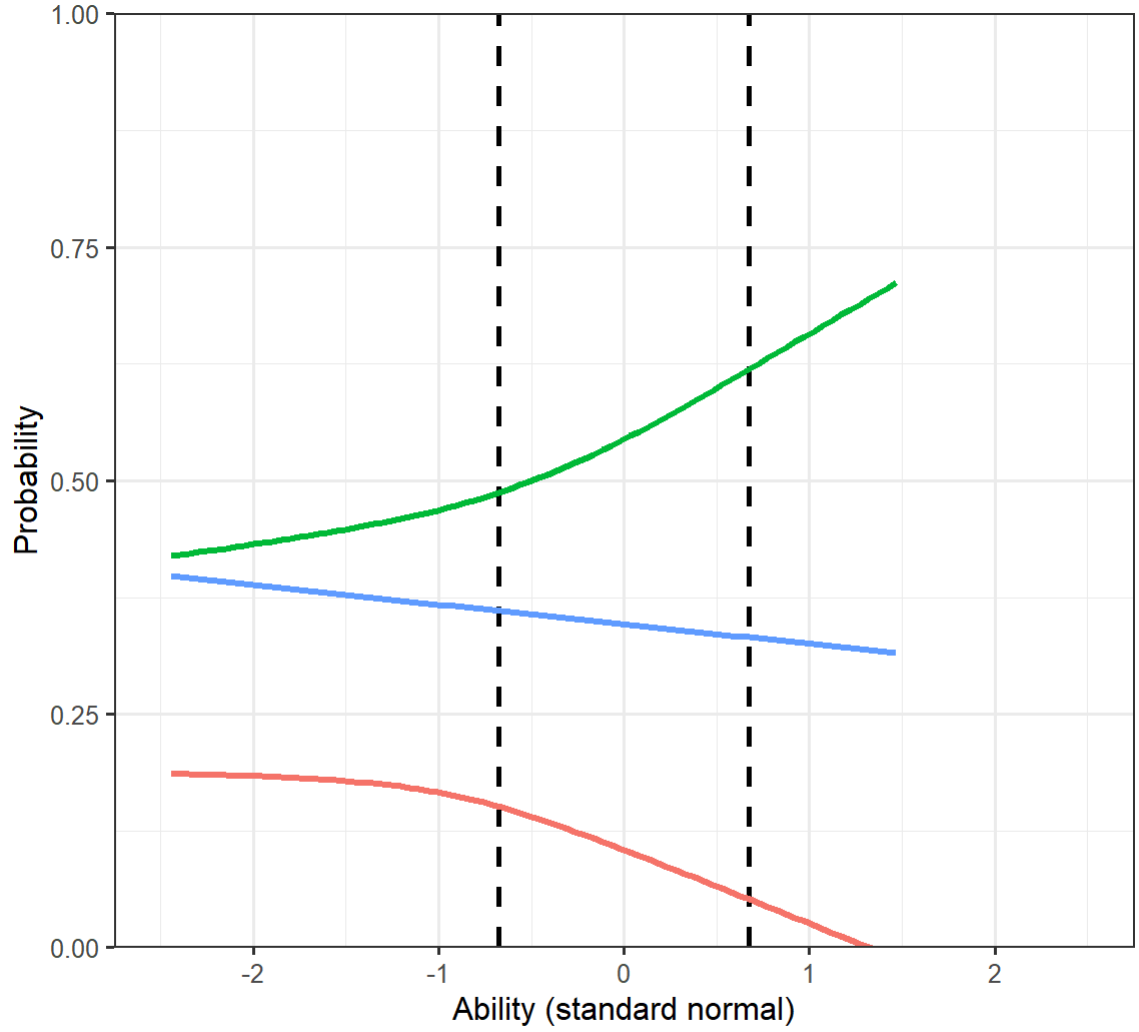
10.2.44.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Drzewa te będą przenosić się do chłodniejszych obszarów, gdzie będą mogły przetrwać.	10	8	0	18	19.6%	10.8%	0.0%	12.9%
Drzewa te będą przystosowywać się do nowych warunków, zmieniając swoje cechy genetyczne.	25	36	12	73	49.0%	48.6%	80.0%	52.1%
Drzewa te będą rosły szybciej i stawały się silniejsze.	16	30	3	49	31.4%	40.5%	20.0%	35.0%
Overall	51	74	15	140	100.0%	100.0%	100.0%	100.0%



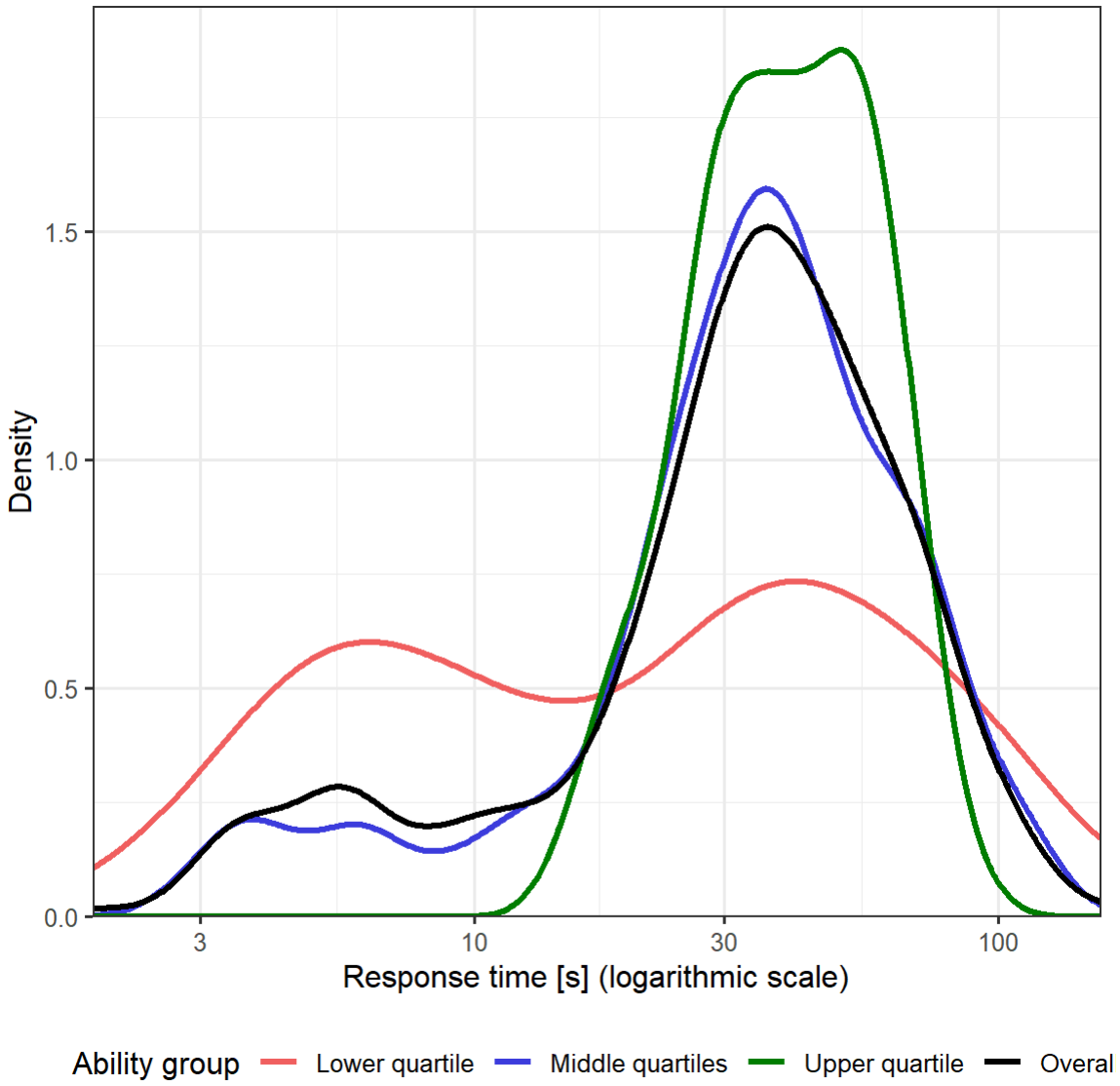
ędą mogły przetrwać. Drzewa te będą przystosowywać się do nowych warunków, zmieniając s\

```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



ędą mogły przetrwać. Drzewa te będą przystosowywać się do nowych warunków, zmieniając s\

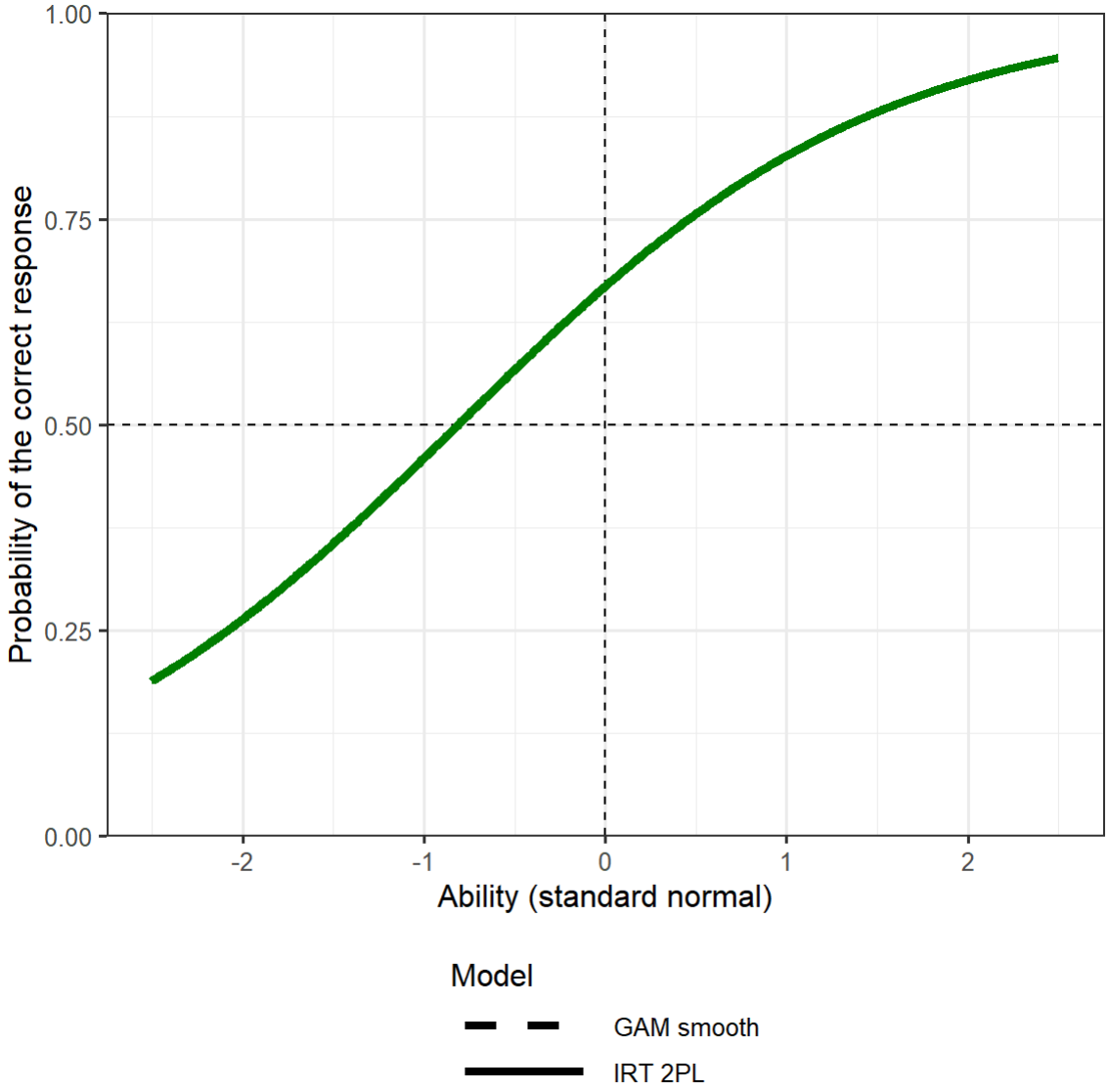
10.2.44.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	1.9	6.8	25.0	46.9	156.4	3.0	1.1
Middle quartiles	2.8	23.6	35.2	49.9	115.3	3.4	0.8
Upper quartile	15.6	29.4	38.5	53.7	85.7	3.7	0.4
Overall	1.9	21.8	34.8	50.4	156.4	3.4	0.8

10.2.45 S890023 (TIMSS)

10.2.45.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	64.89%	319	LLM complex setup	MC	0.35	0.25	0.86	−0.81	0.01

10.2.45.2 Gender DIF:

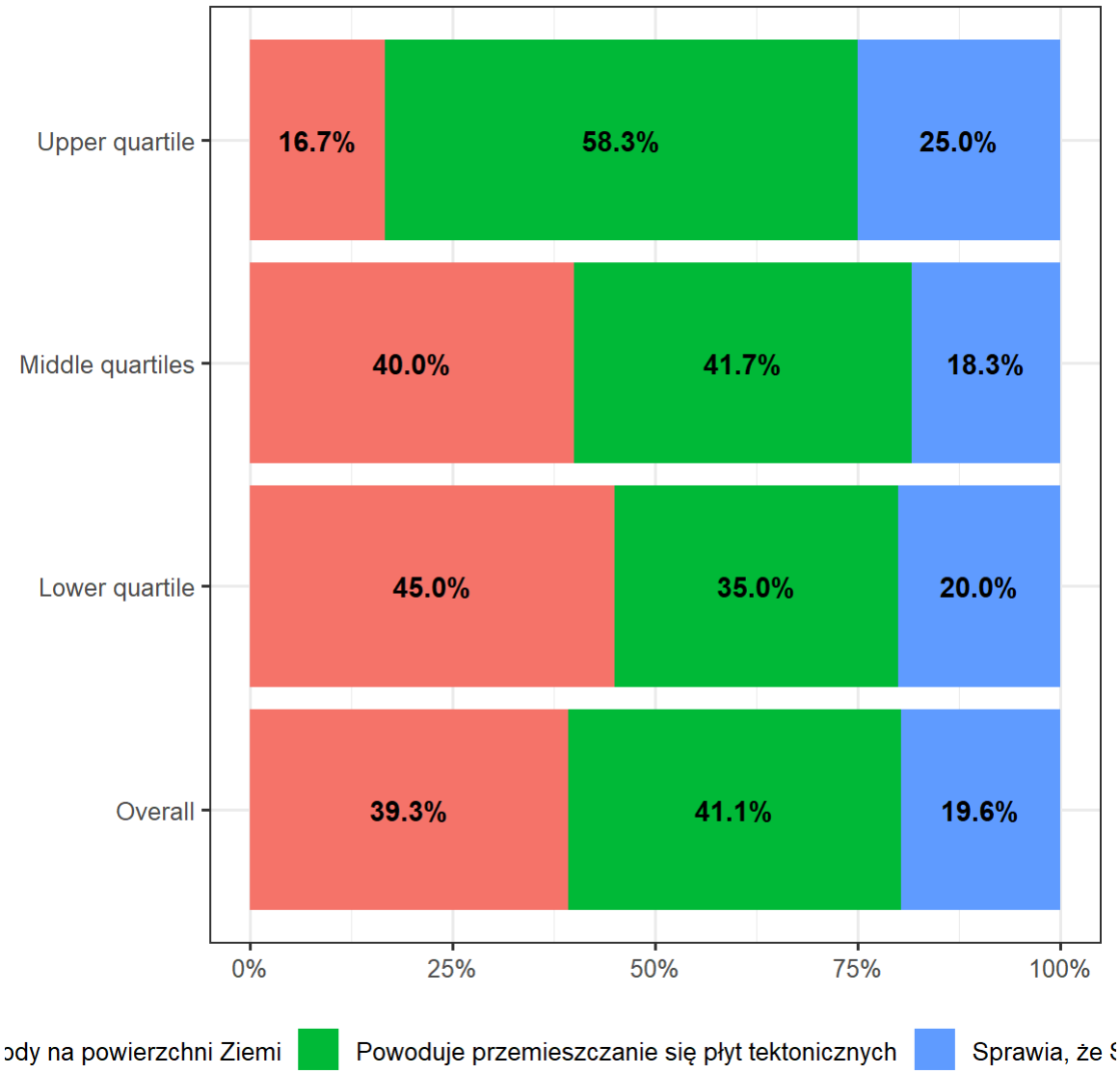
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.37	0.306	1.000	0.007	N/A

10.2.45.3 Frequency of answers:

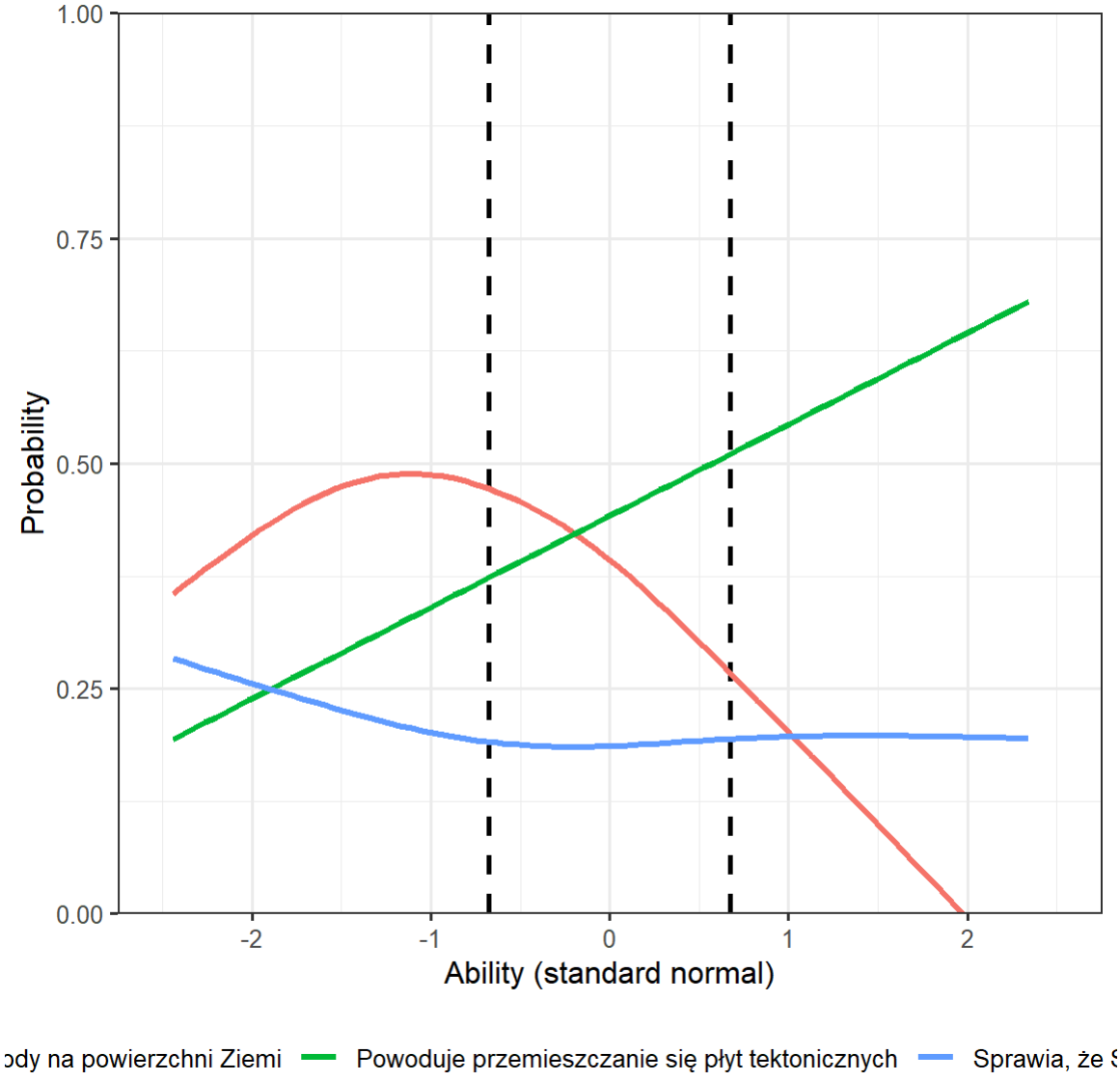
Answer	Frequency	Percent
Chroni Ziemię przed szkodliwym promieniowaniem słonecznym (correct)	207	64.9%
Ogranicza dostęp do wody na powierzchni Ziemi	44	13.8%
Powoduje przemieszczanie się płyt tektonicznych	46	14.4%
Sprawia, że Słońce zawsze świeci w jednym miejscu	22	6.9%
Overall	319	100.0%

10.2.45.4 Distractors functioning:

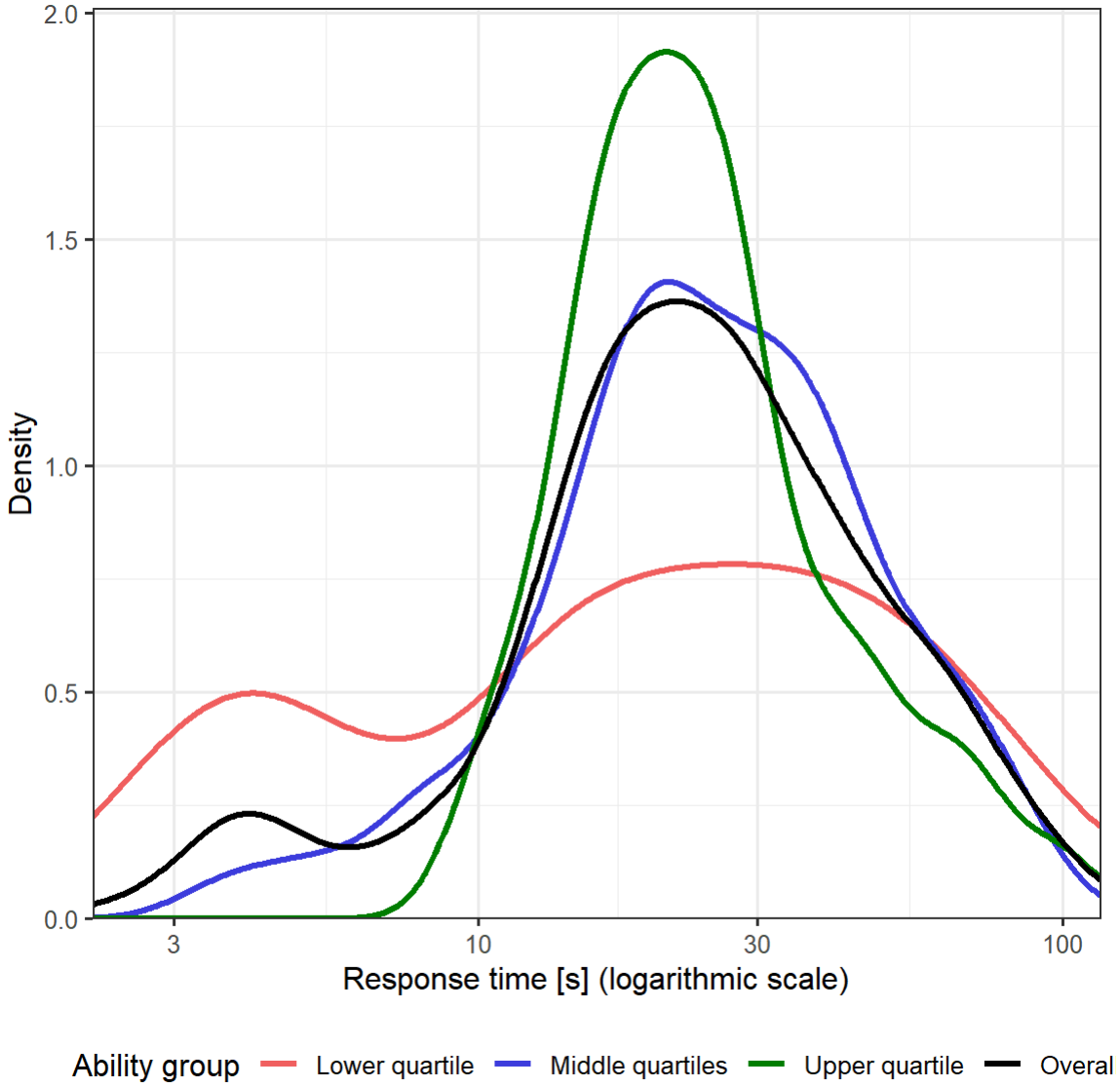
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Ogranicza dostęp do wody na powierzchni Ziemi	18	24	2	44	45.0%	40.0%	16.7%	39.3%
Powoduje przemieszczanie się płyt tektonicznych	14	25	7	46	35.0%	41.7%	58.3%	41.1%
Sprawia, że Słońce zawsze świeci w jednym miejscu	8	11	3	22	20.0%	18.3%	25.0%	19.6%
Overall	40	60	12	112	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



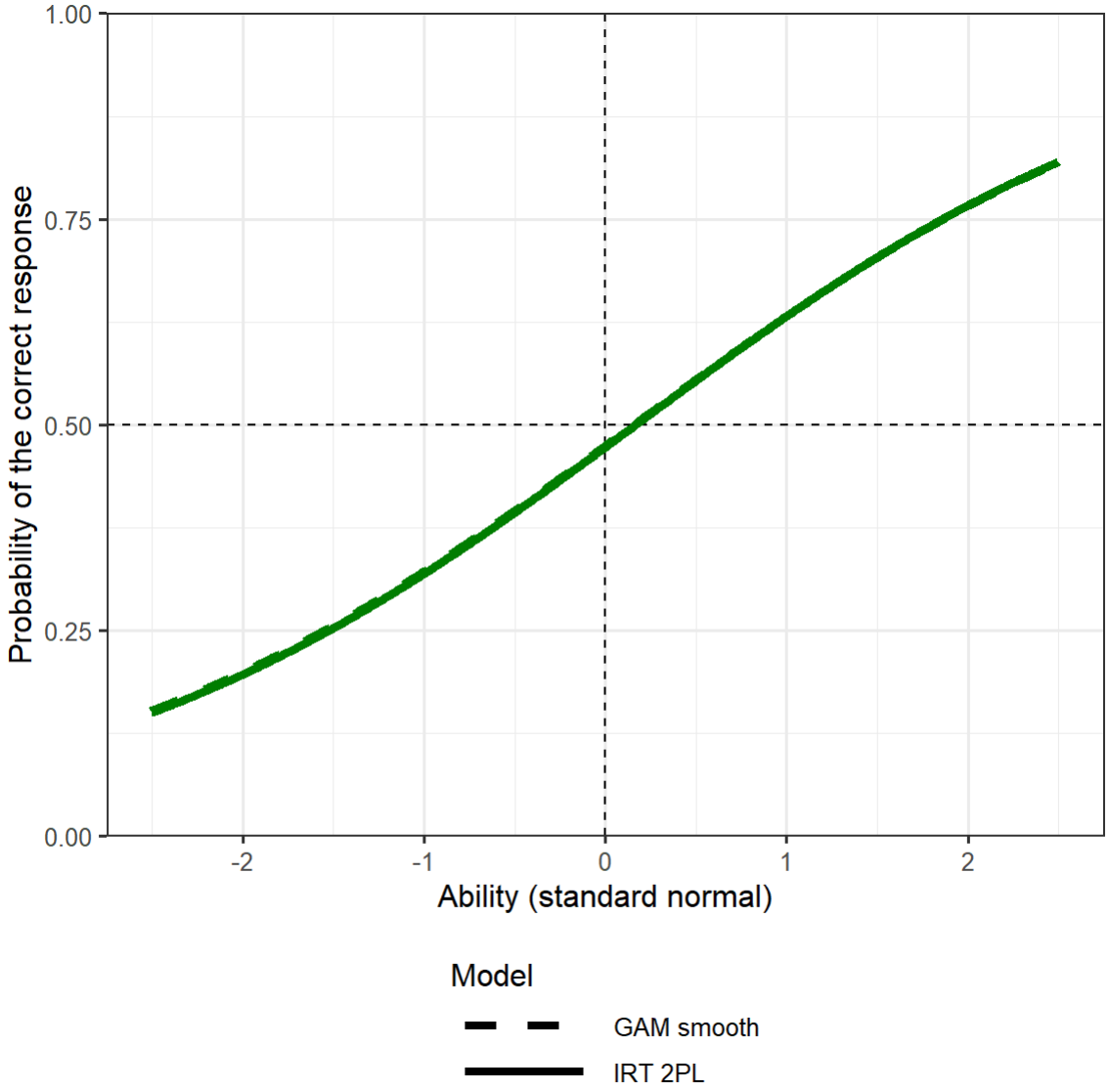
10.2.45.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.2	7.6	18.5	40.7	116.2	2.9	1.1
Middle quartiles	3.5	17.2	24.4	37.7	97.0	3.2	0.7
Upper quartile	10.1	16.9	22.3	30.5	101.8	3.2	0.5
Overall	2.2	15.3	23.5	37.4	116.2	3.1	0.8

10.2.46 S890011 (LLM)

10.2.46.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	47.65%	319	LLM complex setup	MC	0.28	0.18	0.65	0.17	0.00

10.2.46.2 Gender DIF:

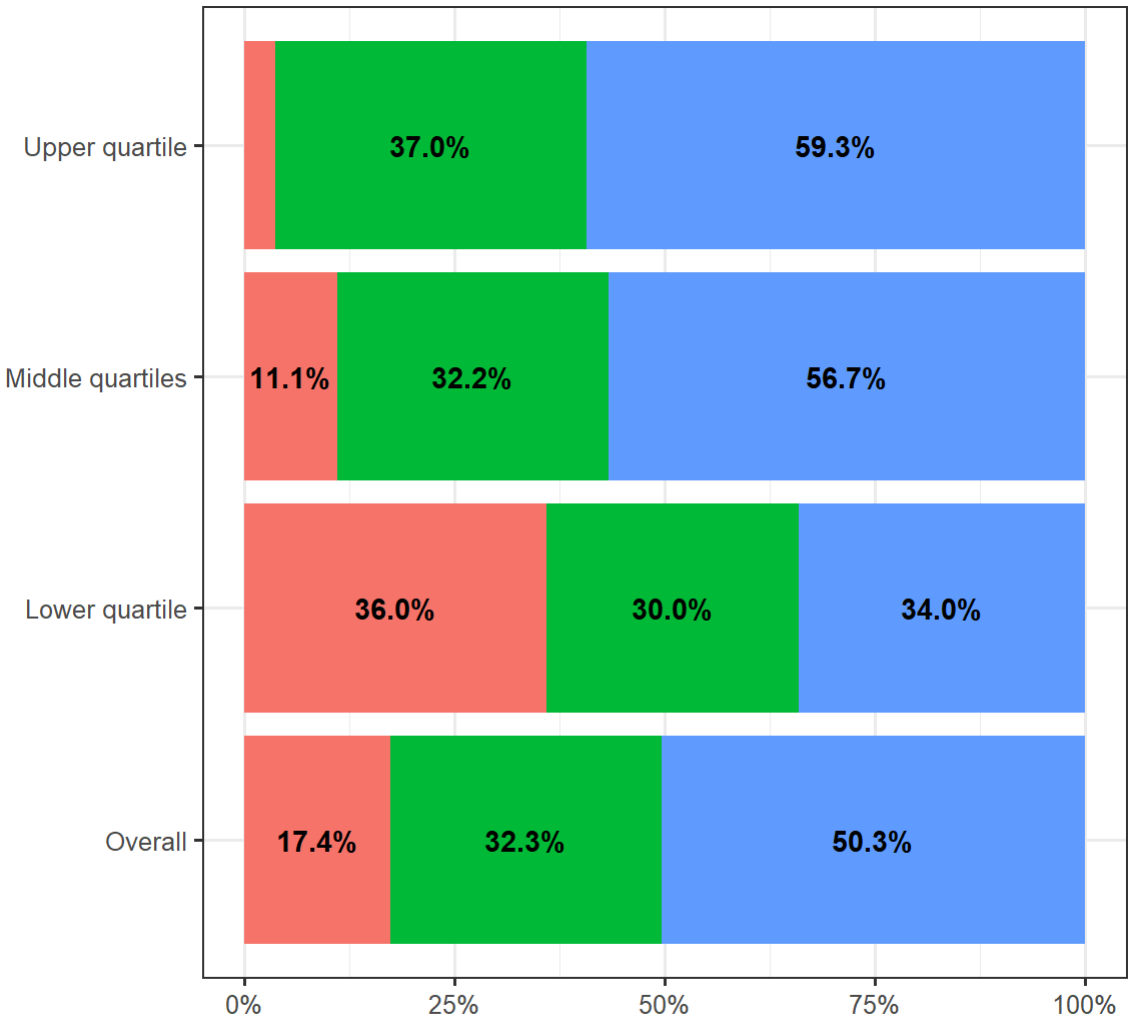
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.25	0.536	1.000	0.003	N/A

10.2.46.3 Frequency of answers:

Answer	Frequency	Percent
Gorący kawałek stanie się jeszcze gorętszy, a zimny jeszcze zimniejszy.	29	9.1%
Gorący kawałek stanie się zimniejszy, a zimny kawałek stanie się cieplejszy. (correct)	152	47.6%
Oba kawałki metalu staną się gorące.	54	16.9%
Temperatura obu kawałków metalu pozostanie bez zmian.	84	26.3%
Overall	319	100.0%

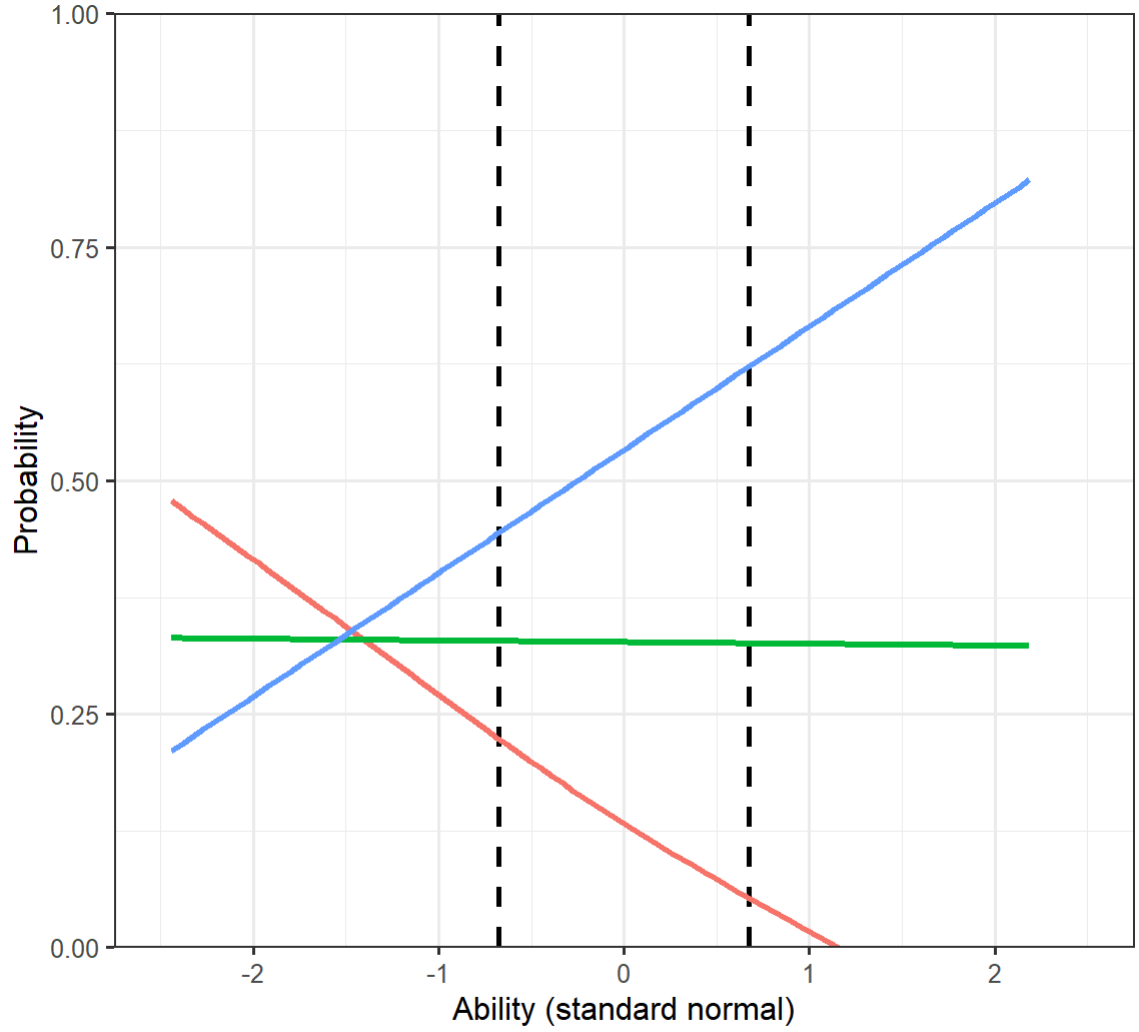
10.2.46.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Gorący kawałek stanie się jeszcze gorętszy, a zimny jeszcze zimniejszy.	18	10	1	29	36.0%	11.1%	3.7%	17.4%
Oba kawałki metalu staną się gorące.	15	29	10	54	30.0%	32.2%	37.0%	32.3%
Temperatura obu kawałków metalu pozostanie bez zmian.	17	51	16	84	34.0%	56.7%	59.3%	50.3%
Overall	50	90	27	167	100.0%	100.0%	100.0%	100.0%



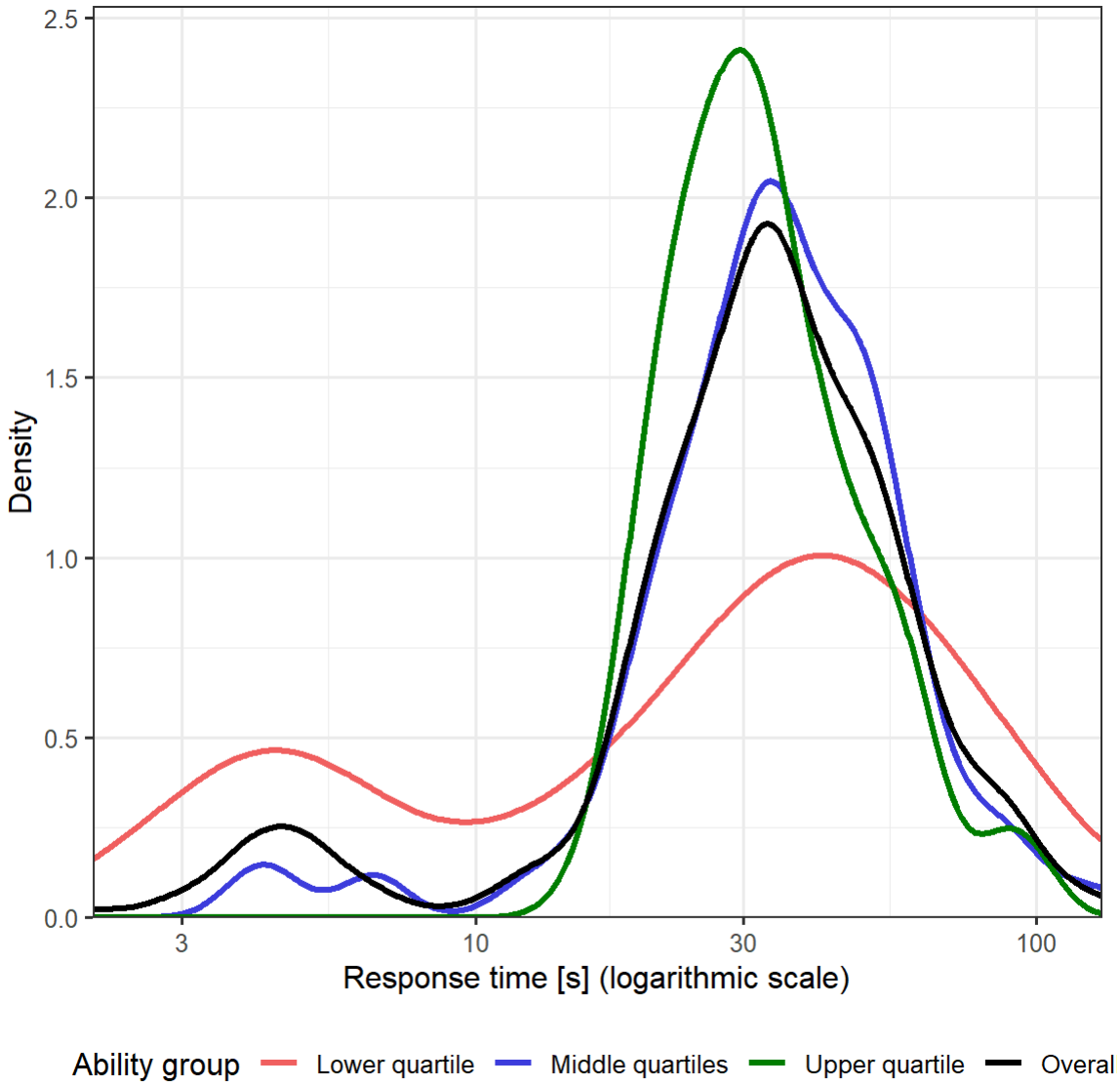
ze gorętszy, a zimny jeszcze zimniejszy. Oba kawałki metalu staną się gorące. Temperatı

```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



ze gorętszy, a zimny jeszcze zimniejszy. Oba kawałki metalu staną się gorące. Temperat

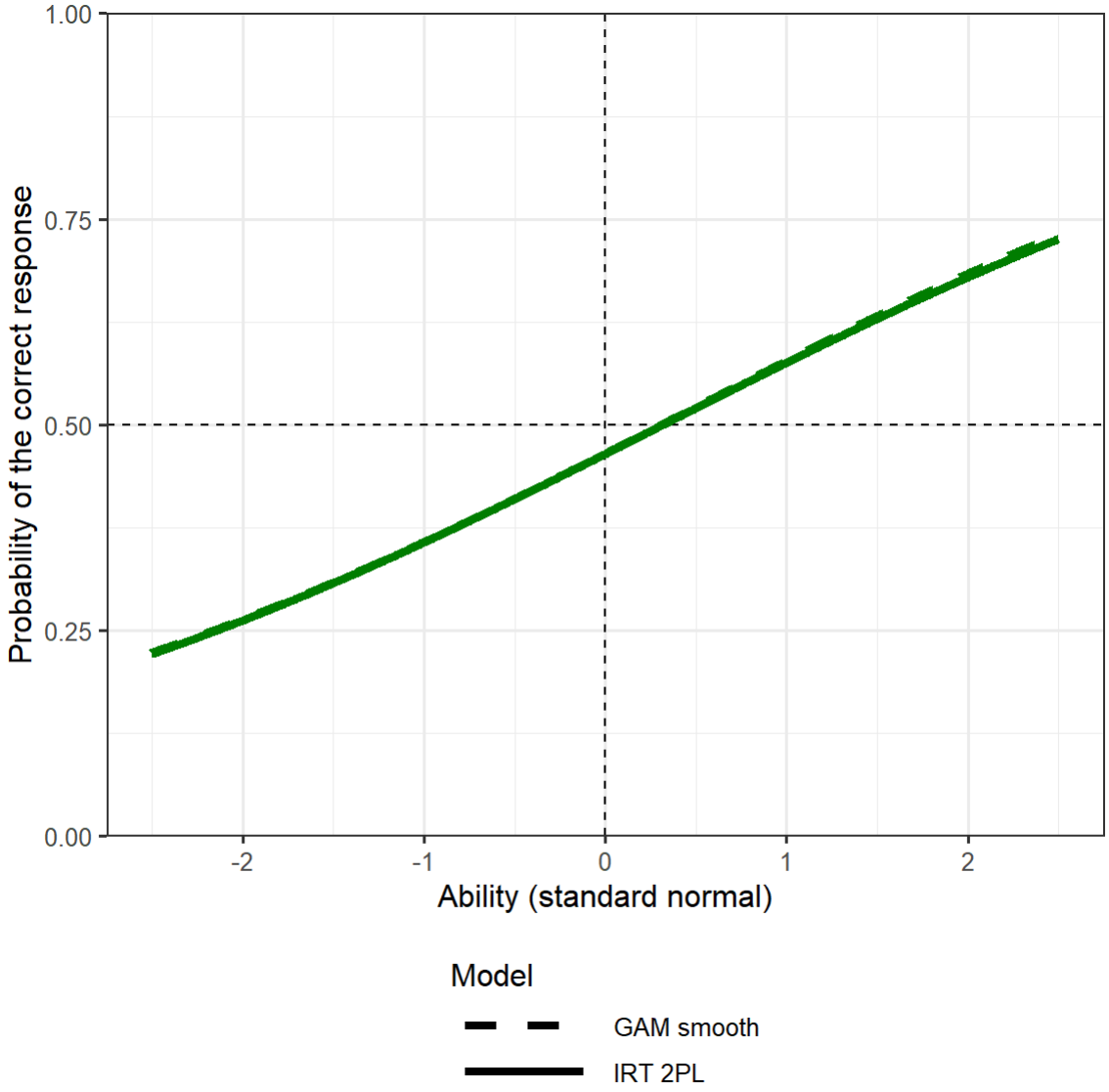
10.2.46.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.1	10.8	33.8	49.5	112.7	3.1	1.1
Middle quartiles	4.0	26.4	34.6	47.0	130.4	3.5	0.6
Upper quartile	16.0	25.2	31.5	41.4	95.8	3.5	0.4
Overall	2.1	23.9	33.4	46.8	130.4	3.4	0.7

10.2.47 S890024 (TIMSS)

10.2.47.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	46.71%	319	LLM complex setup	MC	0.20	0.17	0.45	0.32	0.01

10.2.47.2 Gender DIF:

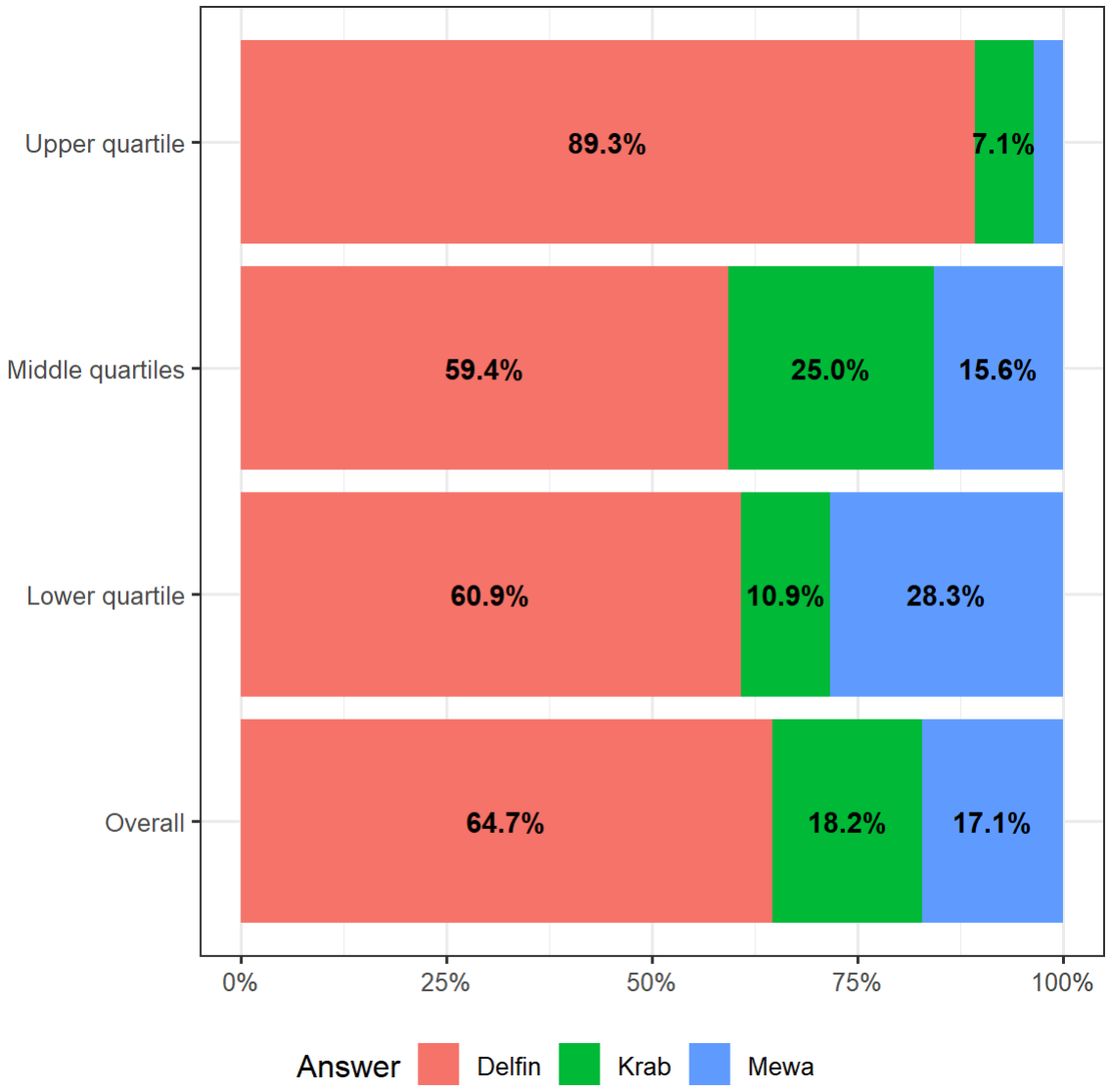
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.04	0.981	1.000	0.000	N/A

10.2.47.3 Frequency of answers:

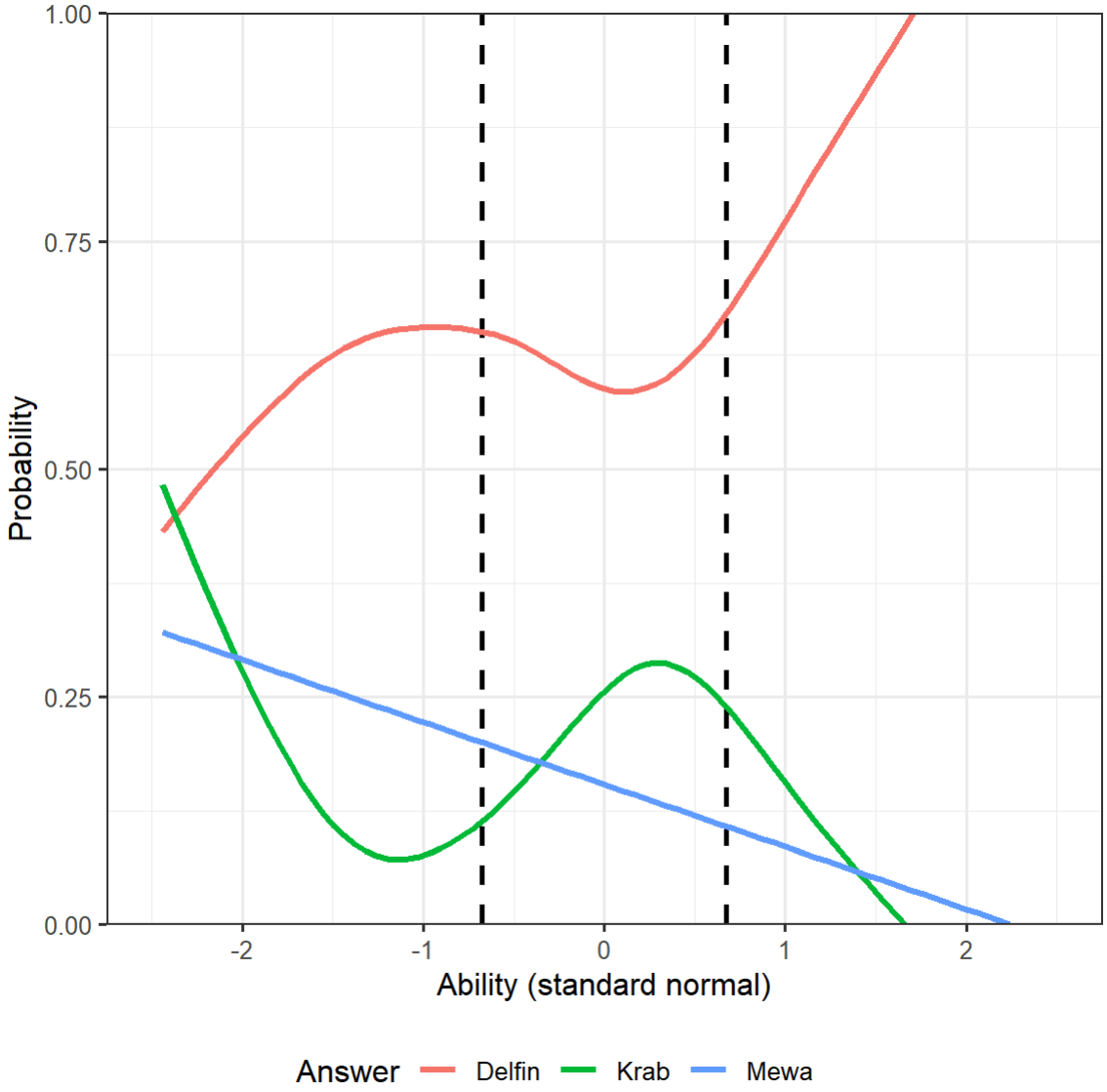
Answer	Frequency	Percent
Delfin	110	34.5%
Karp (correct)	149	46.7%
Krab	31	9.7%
Mewa	29	9.1%
Overall	319	100.0%

10.2.47.4 Distractors functioning:

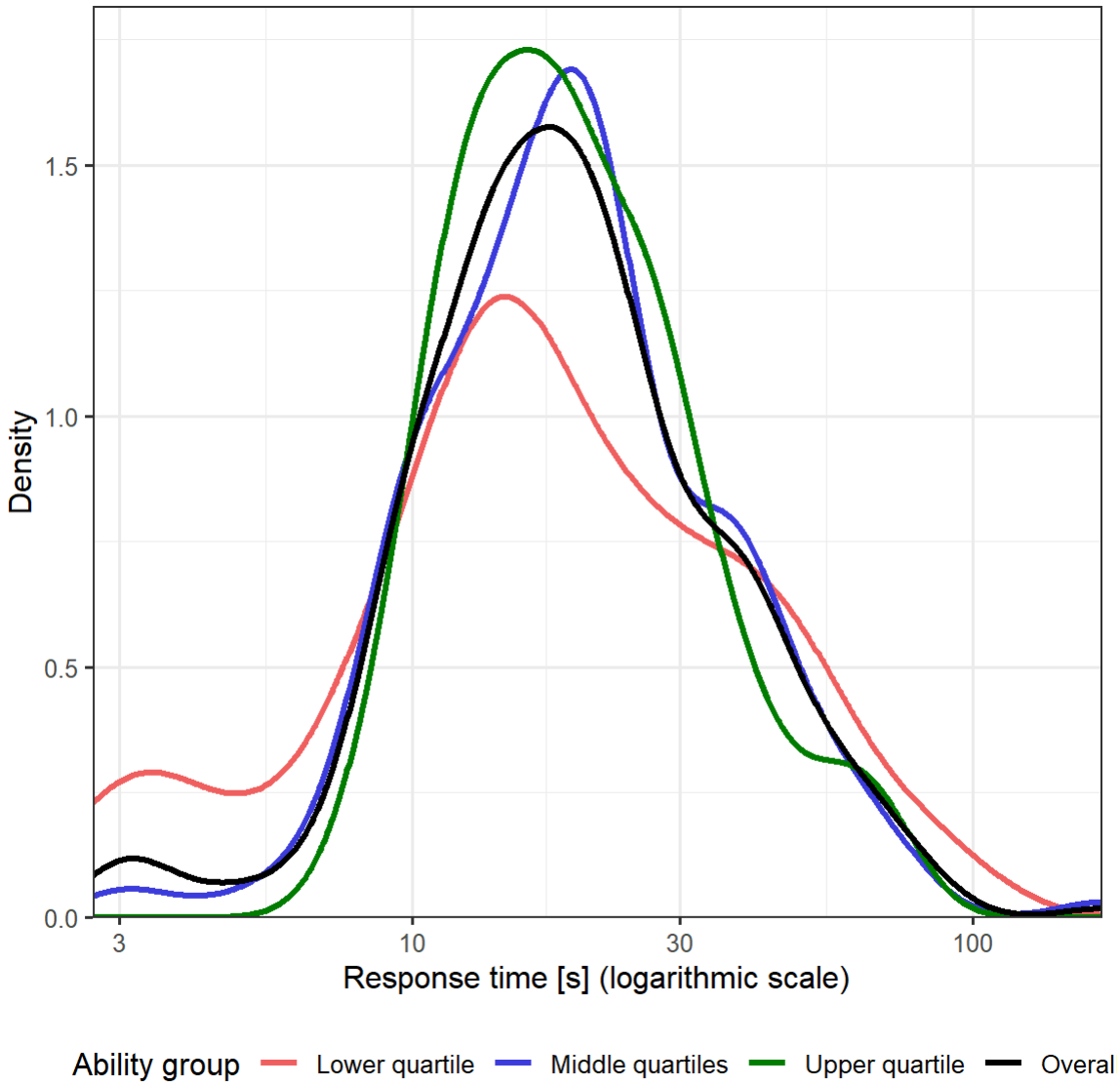
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Delfin	28	57	25	110	60.9%	59.4%	89.3%	64.7%
Krab	5	24	2	31	10.9%	25.0%	7.1%	18.2%
Mewa	13	15	1	29	28.3%	15.6%	3.6%	17.1%
Overall	46	96	28	170	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



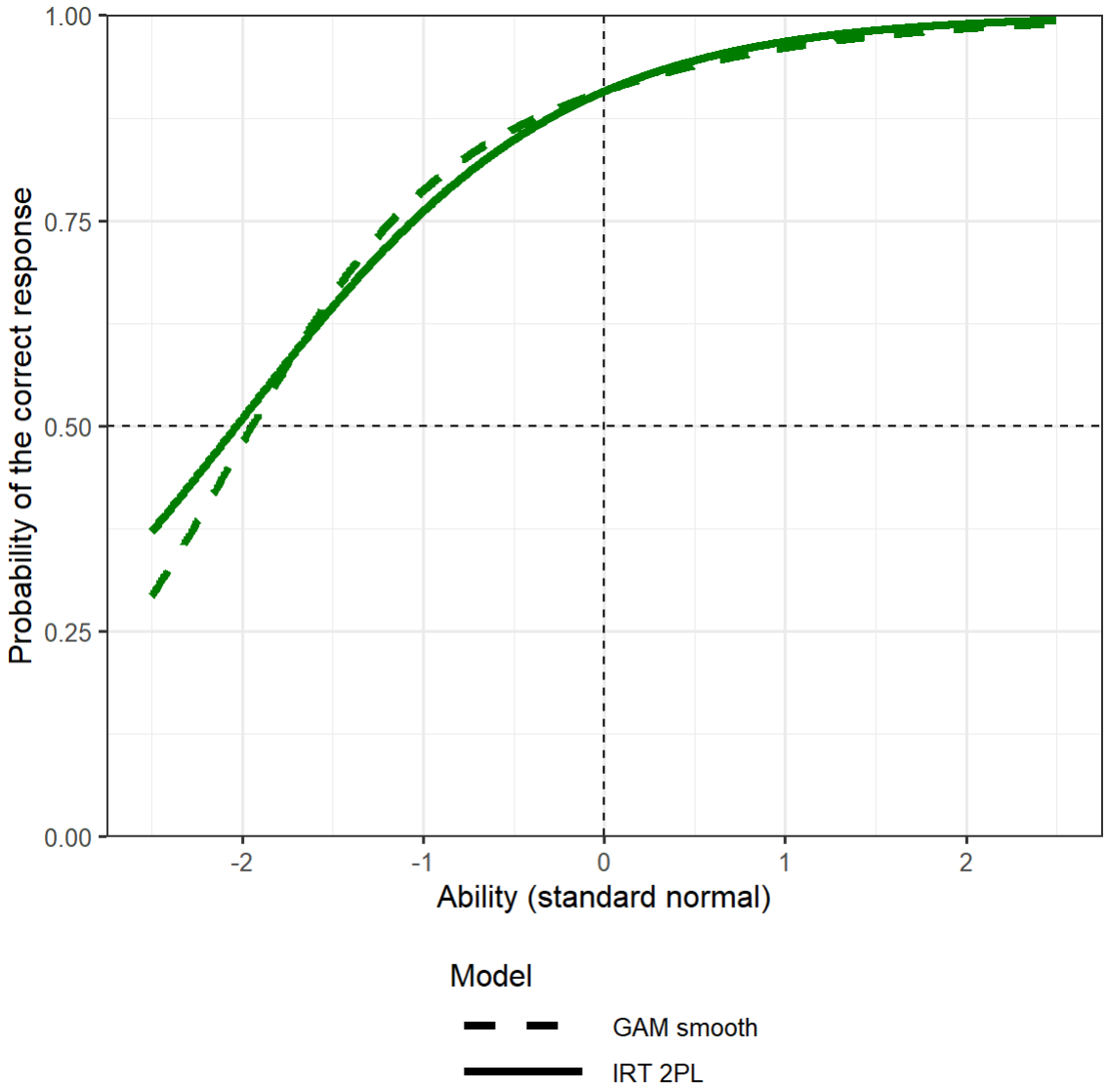
10.2.47.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.7	11.4	15.6	29.9	87.9	2.8	0.8
Middle quartiles	2.9	13.2	19.1	27.9	169.4	3.0	0.6
Upper quartile	7.9	13.5	18.4	26.3	69.5	3.0	0.5
Overall	2.7	12.8	18.2	27.7	169.4	2.9	0.6

10.2.48 S890022 (TIMSS)

10.2.48.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	86.43%	281	LLM complex setup	MC	0.34	0.21	1.12	−2.03	0.00

10.2.48.2 Gender DIF:

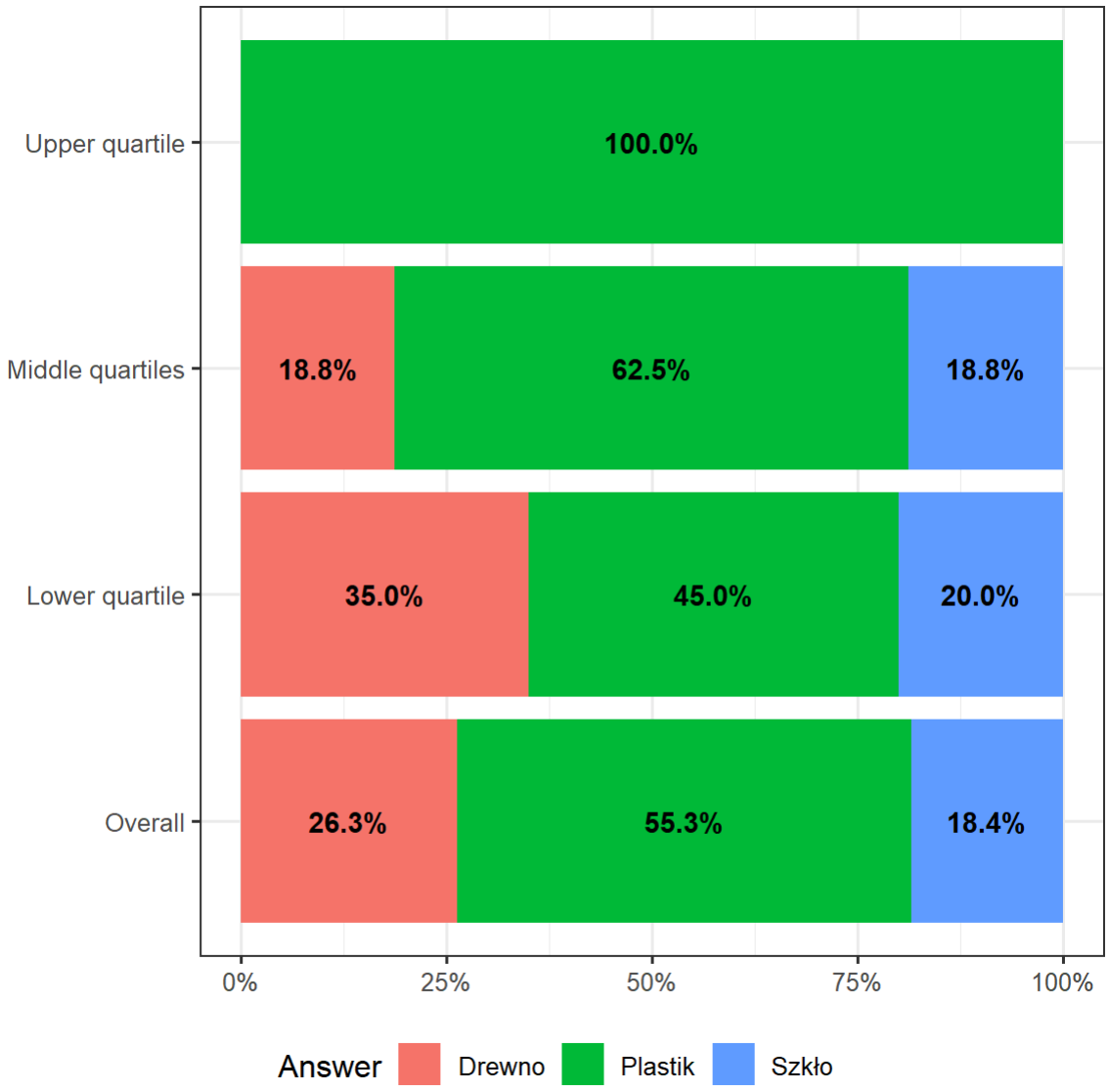
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
4.27	0.118	1.000	0.021	N/A

10.2.48.3 Frequency of answers:

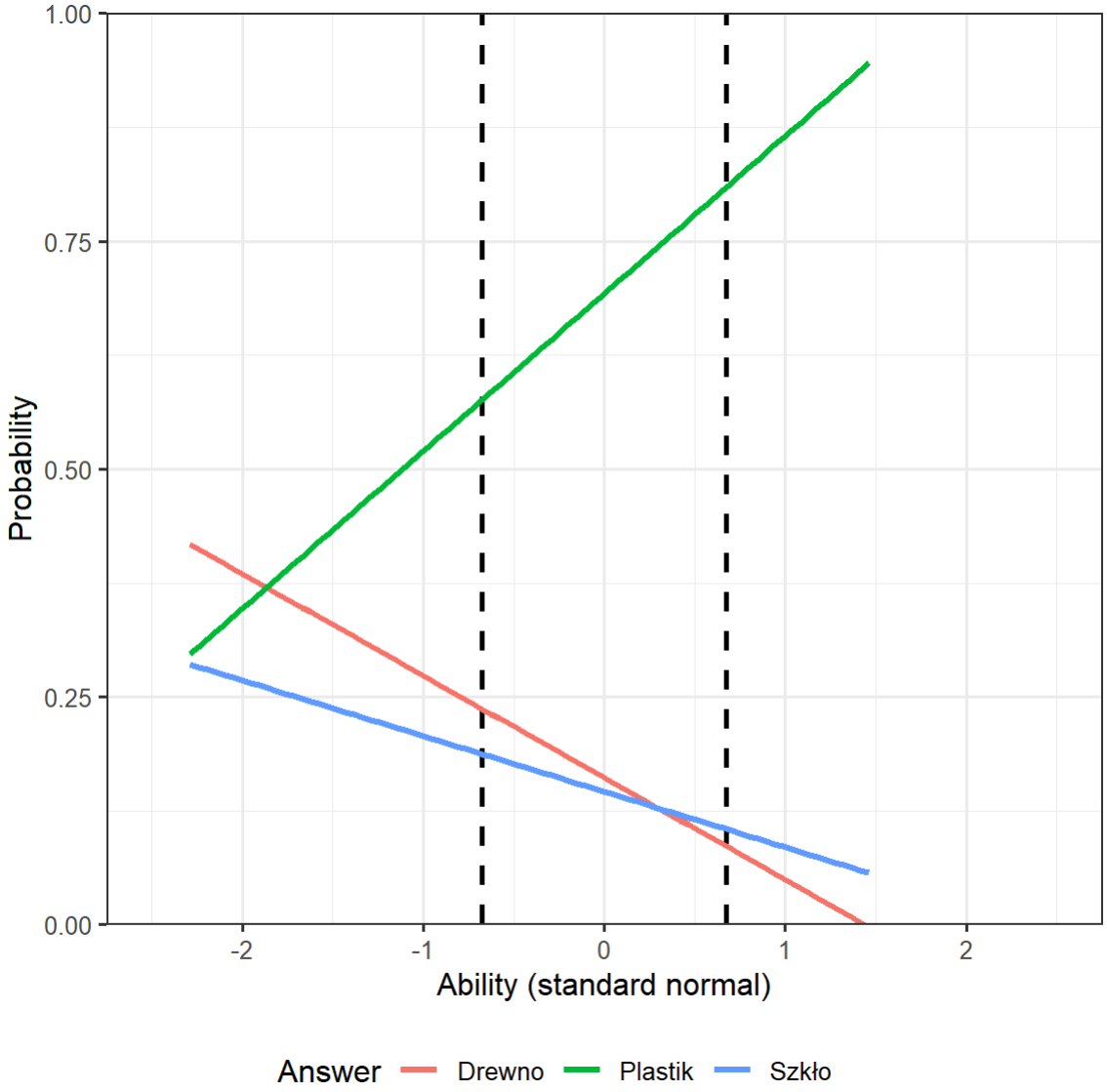
Answer	Frequency	Percent
Drewno	10	3.6%
Metal (correct)	242	86.4%
Plastik	21	7.5%
Szkło	7	2.5%
Overall	280	100.0%

10.2.48.4 Distractors functioning:

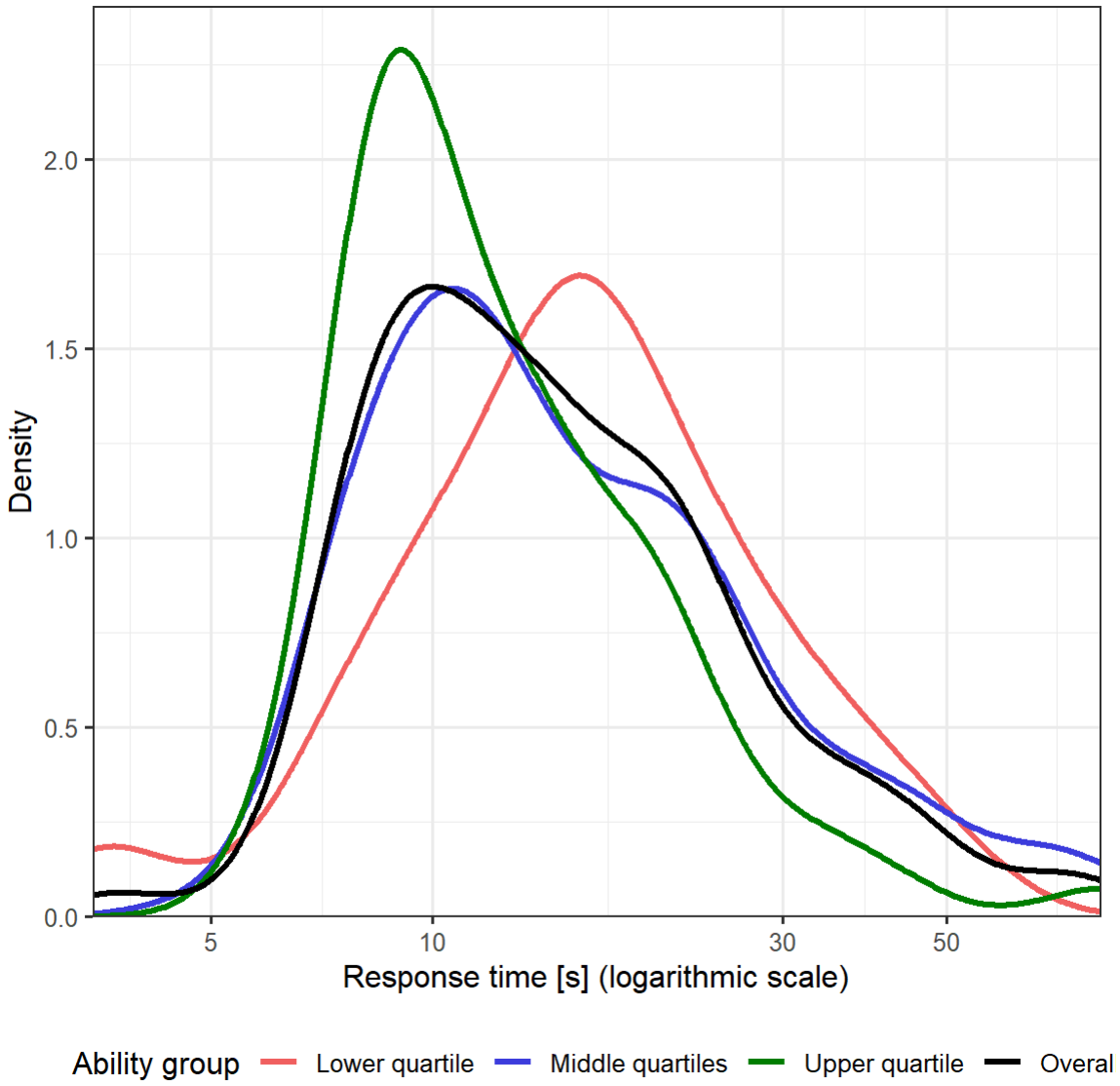
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Drewno	7	3	0	10	35.0%	18.8%	0.0%	26.3%
Plastik	9	10	2	21	45.0%	62.5%	100.0%	55.3%
Szkło	4	3	0	7	20.0%	18.8%	0.0%	18.4%
Overall	20	16	2	38	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



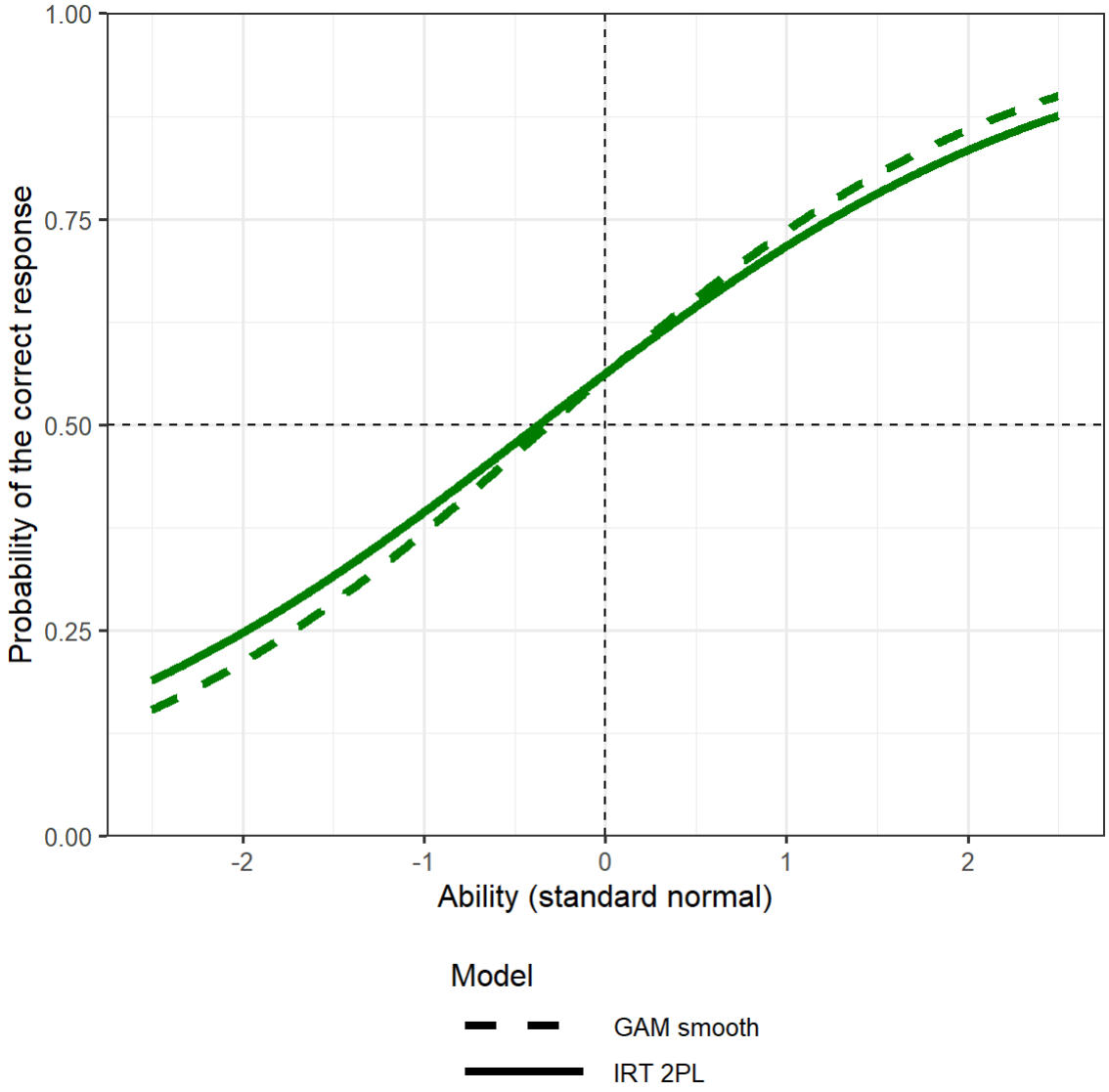
10.2.48.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.4	10.9	15.7	22.7	53.3	2.8	0.6
Middle quartiles	4.9	9.6	13.3	21.8	79.1	2.7	0.6
Upper quartile	6.0	8.5	10.8	15.9	81.0	2.5	0.5
Overall	3.4	9.4	13.4	20.8	81.0	2.7	0.6

10.2.49 S051121 (TIMSS)

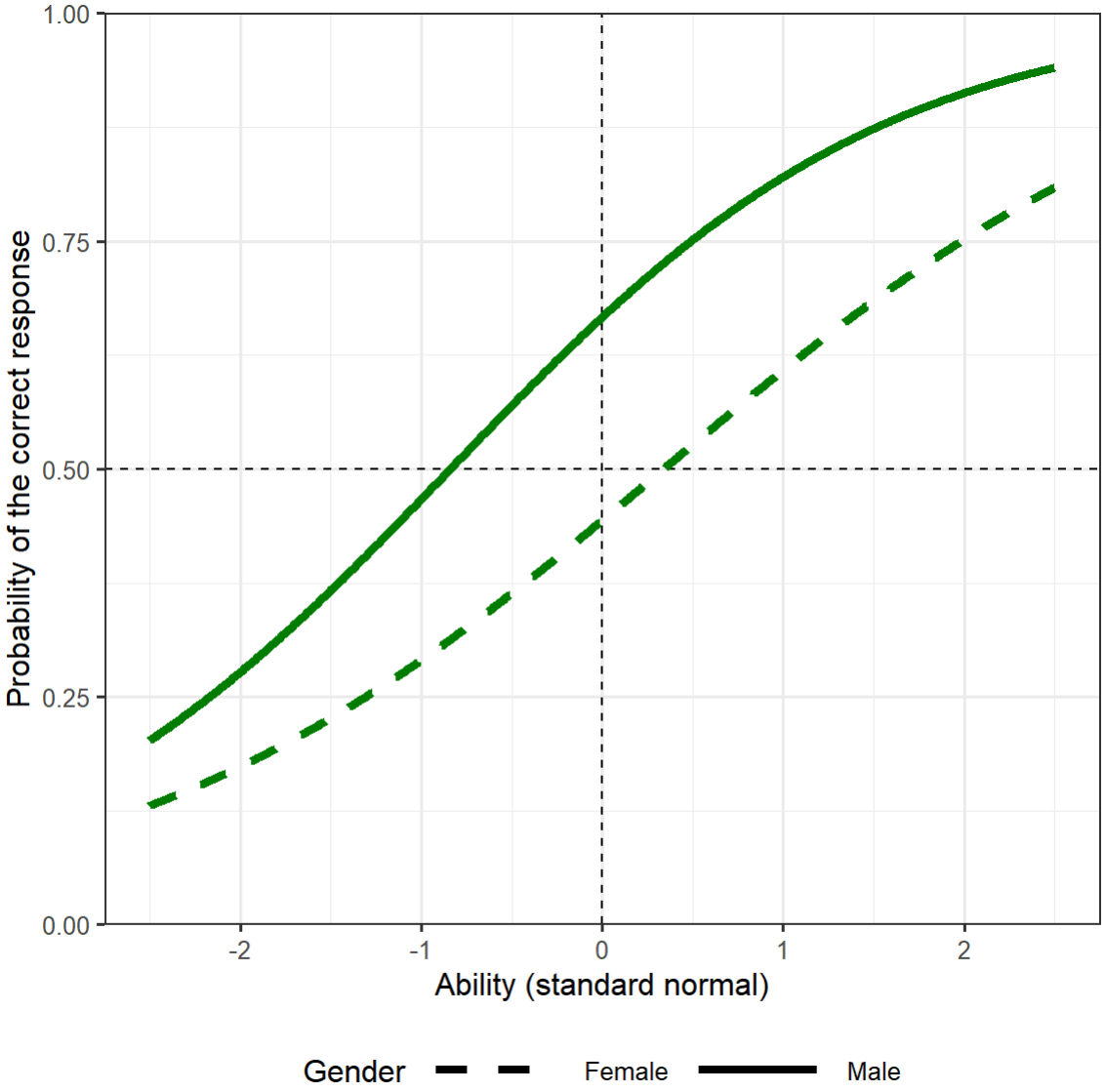
10.2.49.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	56.46%	294	TIMSS	MC	0.34	0.29	0.68	−0.37	0.003

10.2.49.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
17.64	0.000	0.014	0.051	moderate



10.2.49.3 Frequency of answers:

Answer	Frequency	Percent
Nie, Nie, Nie, Nie, Nie	13	4.4%
Nie, Nie, Nie, Nie, Tak	11	3.7%
Nie, Nie, Nie, Tak, Tak	1	0.3%
Nie, Nie, Tak, Nie, Nie	8	2.7%
Nie, Nie, Tak, Nie, Tak (correct)	166	56.5%
Nie, Nie, Tak, Tak, Nie	1	0.3%
Nie, Nie, Tak, Tak, Tak	16	5.4%
Nie, Tak, Nie, Nie, Tak	2	0.7%
Nie, Tak, Nie, Tak, Nie	12	4.1%
Nie, Tak, Tak, Nie, Nie	2	0.7%

Answer	Frequency	Percent
Nie, Tak, Tak, Nie, Tak	23	7.8%
Nie, Tak, Tak, Tak, Nie	2	0.7%
Nie, Tak, Tak, Tak, Tak	4	1.4%
Tak, Nie, Nie, Nie, Nie	1	0.3%
Tak, Nie, Nie, Nie, Tak	1	0.3%
Tak, Nie, Nie, Tak, Nie	4	1.4%
Tak, Nie, Tak, Nie, Nie	1	0.3%
Tak, Nie, Tak, Nie, Tak	7	2.4%
Tak, Nie, Tak, Tak, Nie	3	1.0%
Tak, Nie, Tak, Tak, Tak	1	0.3%
Tak, Tak, Nie, Nie, Tak	4	1.4%
Tak, Tak, Tak, Nie, Nie	1	0.3%
Tak, Tak, Tak, Nie, Tak	4	1.4%
Tak, Tak, Tak, Tak, Tak	6	2.0%
Overall	294	100.0%

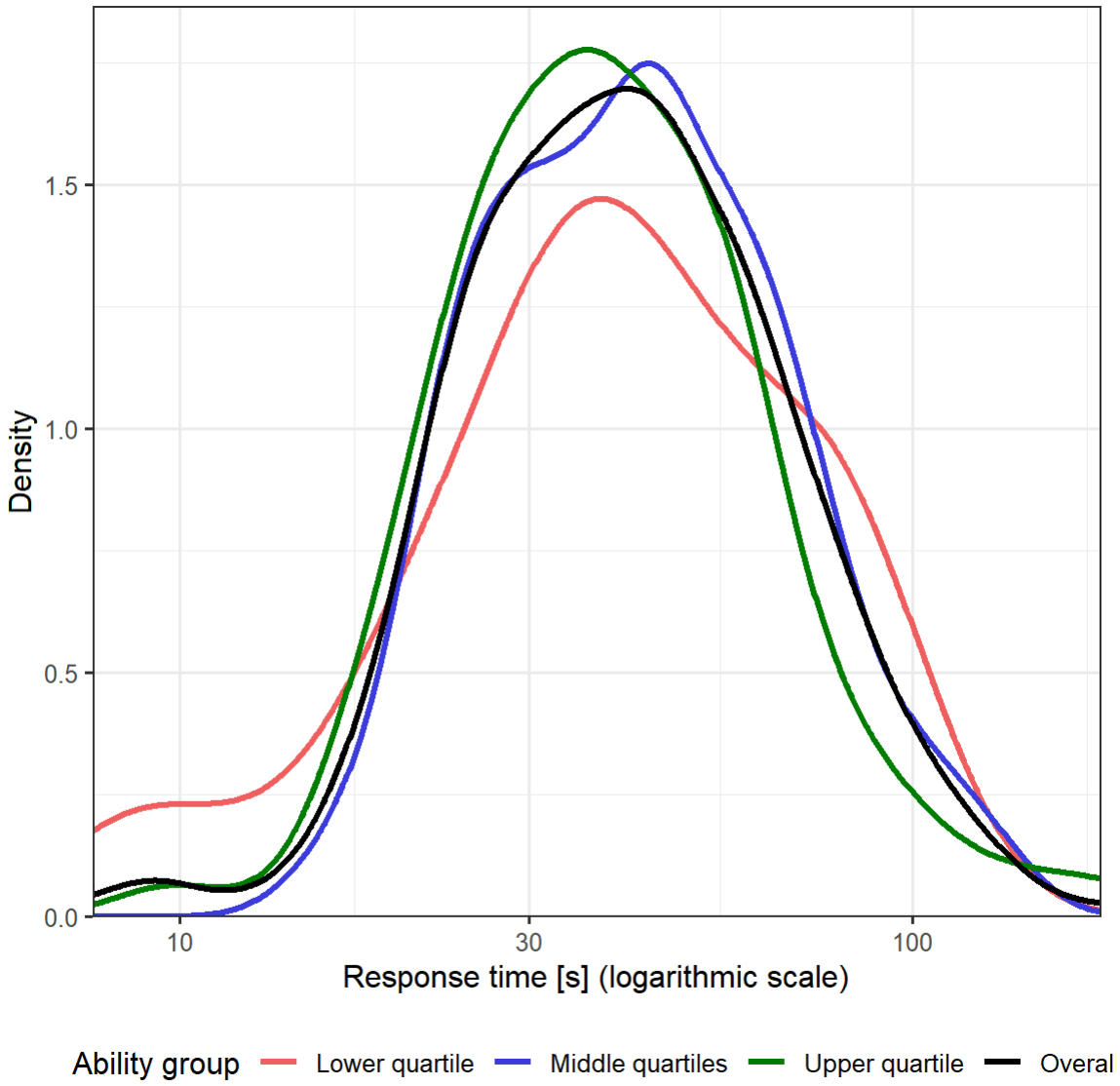
10.2.49.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Nie, Nie, Nie, Nie, Nie	5	8	0	13	13.2%	11.0%	0.0%	10.2%
Nie, Nie, Nie, Nie, Tak	3	6	2	11	7.9%	8.2%	11.8%	8.6%
Nie, Nie, Tak, Nie, Nie	2	6	0	8	5.3%	8.2%	0.0%	6.2%

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Nie, Nie, Tak, Tak, Tak	2	10	4	16	5.3%	13.7%	23.5%	12.5%
Nie, Tak, Nie, Nie, Tak	1	1	0	2	2.6%	1.4%	0.0%	1.6%
Nie, Tak, Nie, Tak, Nie	2	7	3	12	5.3%	9.6%	17.6%	9.4%
Nie, Tak, Tak, Nie, Tak	4	16	3	23	10.5%	21.9%	17.6%	18.0%
Nie, Tak, Tak, Tak, Nie	1	0	1	2	2.6%	0.0%	5.9%	1.6%
Nie, Tak, Tak, Tak, Tak	1	2	1	4	2.6%	2.7%	5.9%	3.1%
Tak, Nie, Nie, Tak, Nie	2	2	0	4	5.3%	2.7%	0.0%	3.1%
Tak, Nie, Tak, Nie, Tak	2	5	0	7	5.3%	6.8%	0.0%	5.5%
Tak, Nie, Tak, Tak, Nie	1	2	0	3	2.6%	2.7%	0.0%	2.3%
Tak, Nie, Tak, Tak, Tak	1	0	0	1	2.6%	0.0%	0.0%	0.8%
Tak, Tak, Nie, Nie, Tak	3	1	0	4	7.9%	1.4%	0.0%	3.1%

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Tak, Tak, Tak, Nie, Tak	3	1	0	4	7.9%	1.4%	0.0%	3.1%
Tak, Tak, Tak, Tak, Tak	5	1	0	6	13.2%	1.4%	0.0%	4.7%
Overall	38	73	17	128	100.0%	100.0%	100.0%	100.0%
Nie, Nie, Nie, Tak, Tak	0	1	0	1	0.0%	1.4%	0.0%	0.8%
Nie, Tak, Tak, Nie, Nie	0	1	1	2	0.0%	1.4%	5.9%	1.6%
Tak, Nie, Nie, Nie, Nie	0	1	0	1	0.0%	1.4%	0.0%	0.8%
Tak, Nie, Nie, Nie, Tak	0	1	0	1	0.0%	1.4%	0.0%	0.8%
Tak, Tak, Tak, Nie, Nie	0	1	0	1	0.0%	1.4%	0.0%	0.8%
Nie, Nie, Tak, Tak, Nie	0	0	1	1	0.0%	0.0%	5.9%	0.8%
Tak, Nie, Tak, Nie, Nie	0	0	1	1	0.0%	0.0%	5.9%	0.8%

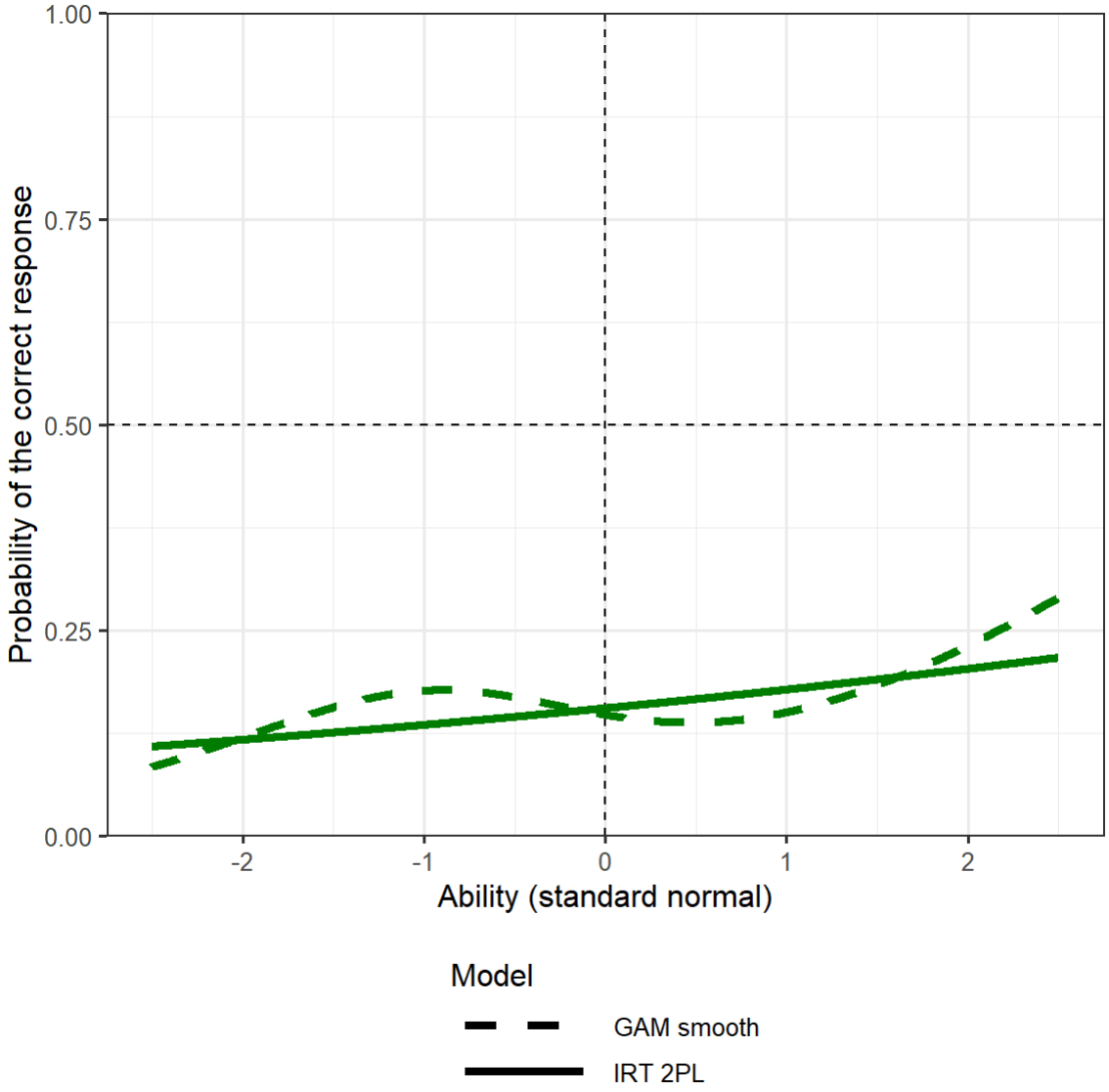
10.2.49.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	7.6	27.6	38.7	60.3	107.8	3.6	0.6
Middle quartiles	15.6	29.1	42.9	58.2	131.7	3.7	0.5
Upper quartile	9.8	26.1	36.6	52.7	180.8	3.6	0.5
Overall	7.6	28.2	41.1	57.2	180.8	3.7	0.5

10.2.50 S100014 (TIMSS)

10.2.50.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	15.67%	319	LLM simple setup	MC	0.02	0.04	0.16	10.35	0.015

10.2.50.2 Gender DIF:

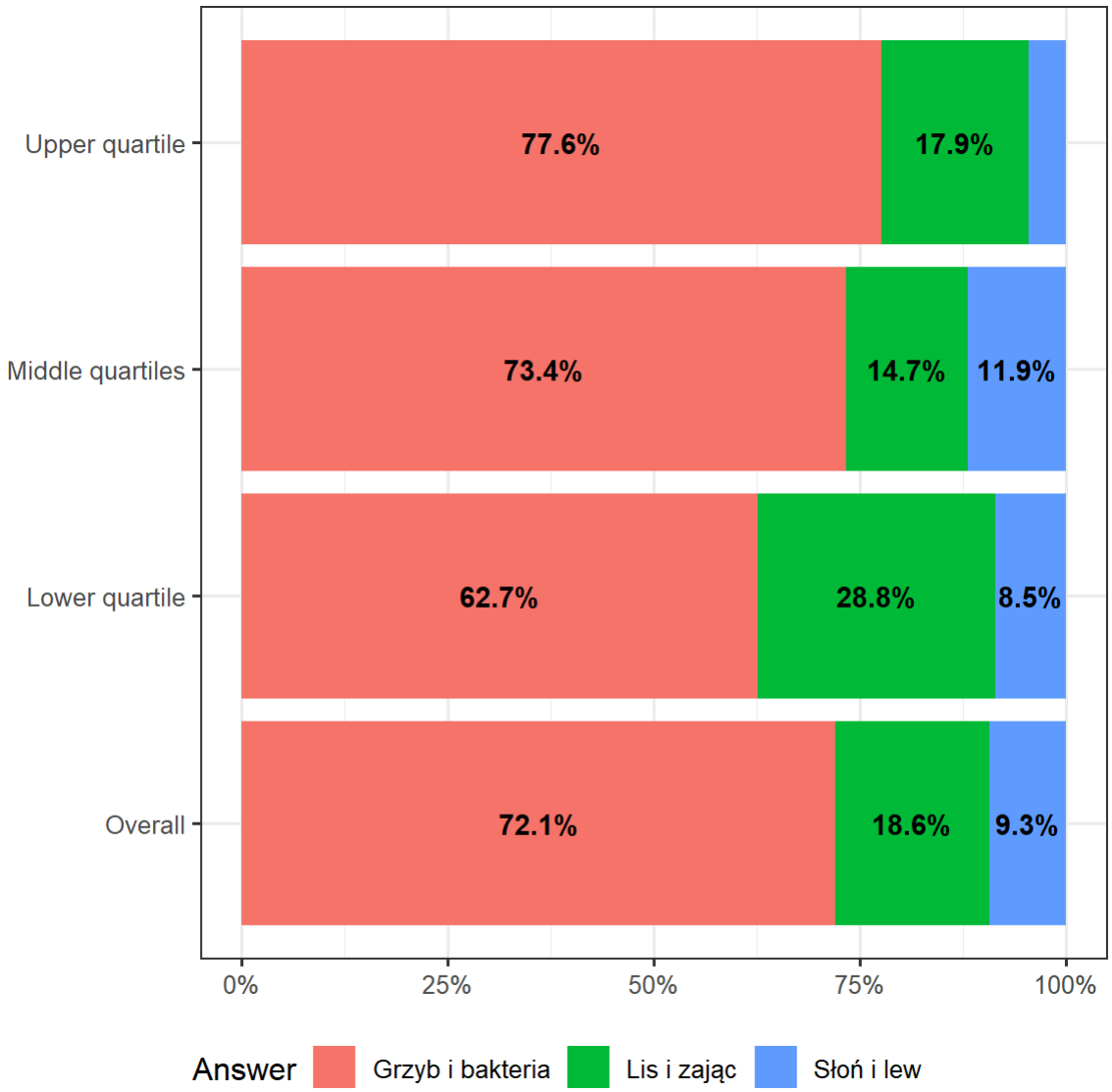
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
6.18	0.045	1.000	0.025	N/A

10.2.50.3 Frequency of answers:

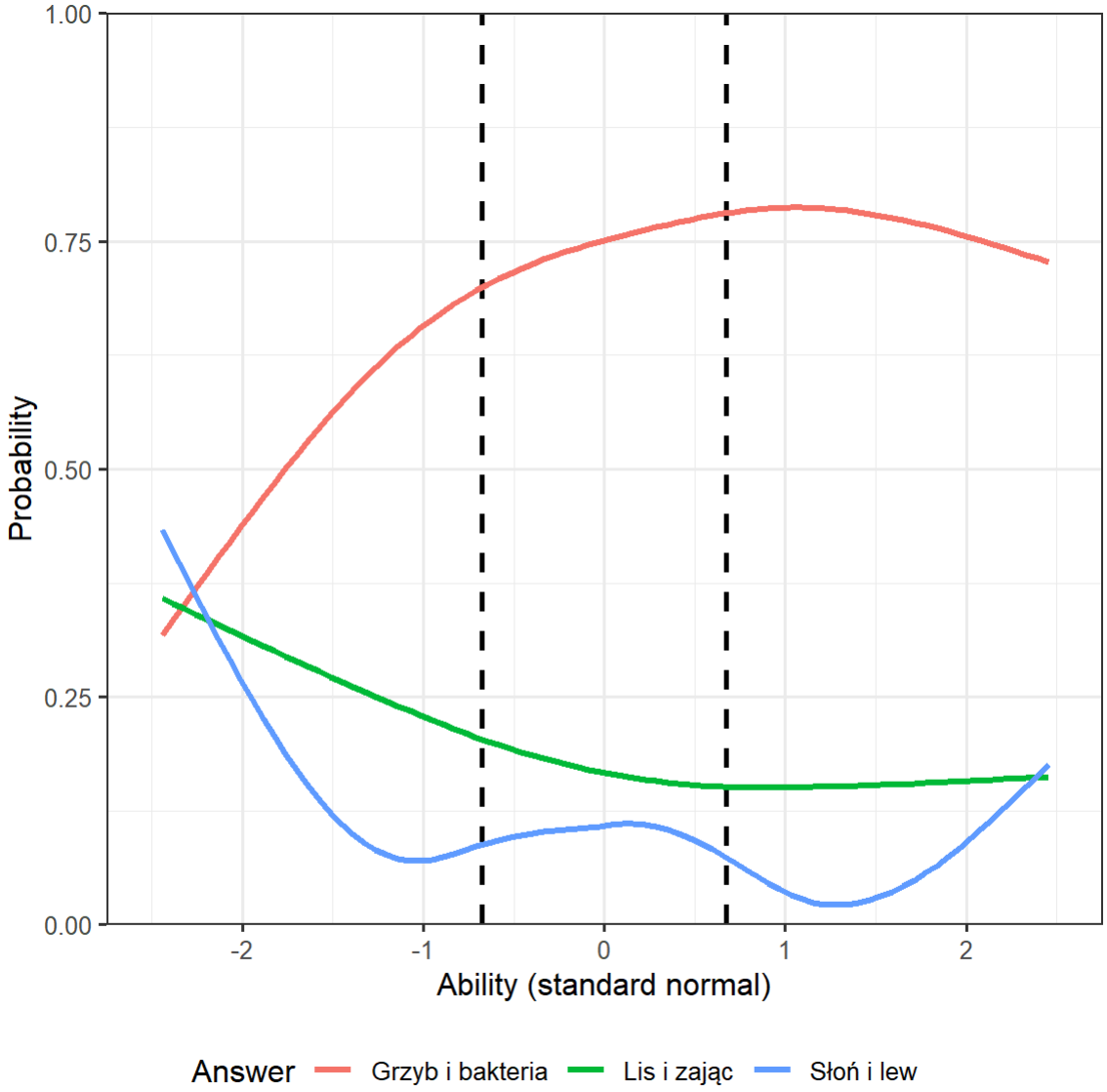
Answer	Frequency	Percent
Grzyb i bakteria	194	60.8%
Lis i zając	50	15.7%
Rekin i remora (ryba podczepka) (correct)	50	15.7%
Słoń i lew	25	7.8%
Overall	319	100.0%

10.2.50.4 Distractors functioning:

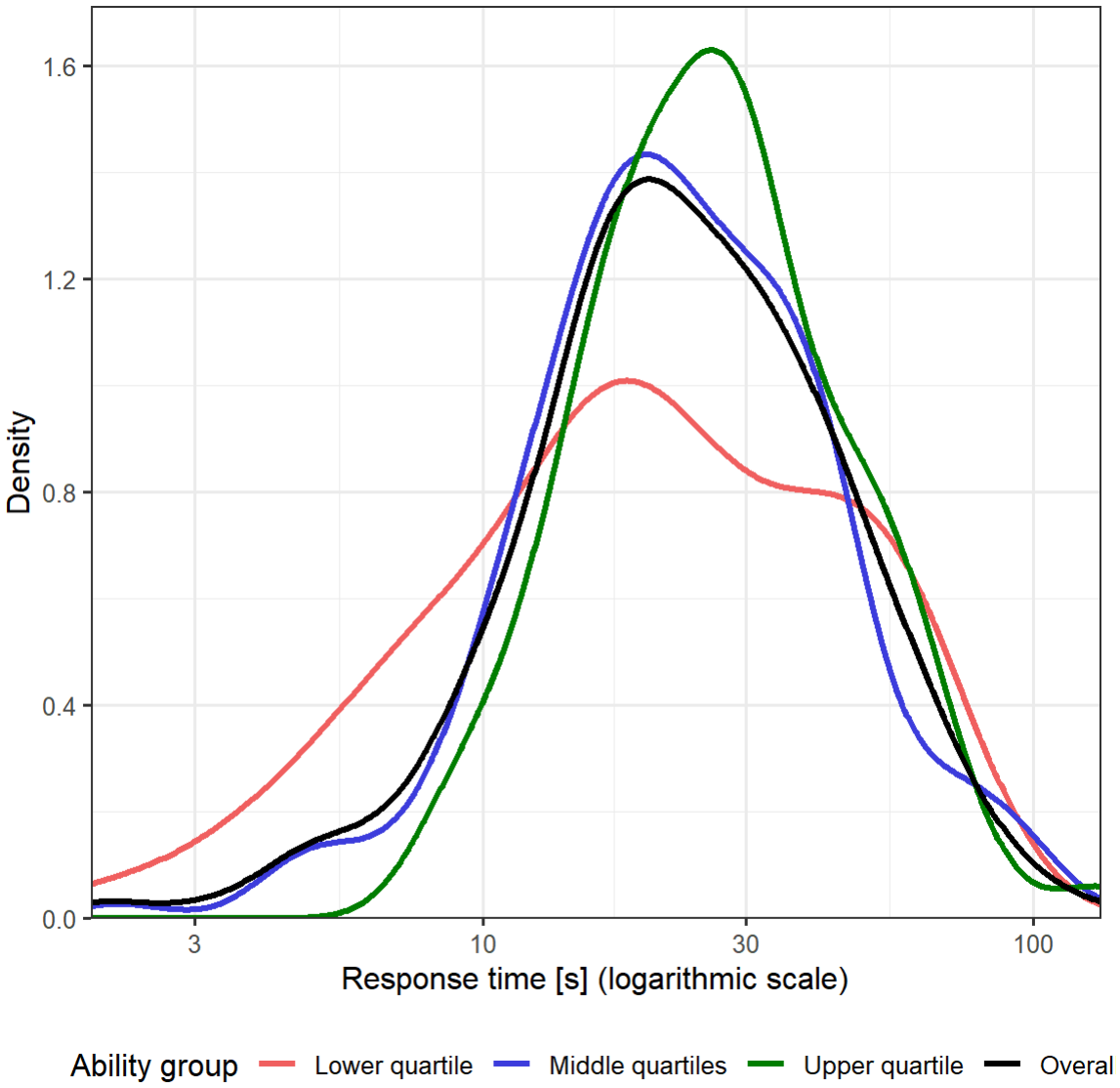
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Grzyb i bakteria	37	105	52	194	62.7%	73.4%	77.6%	72.1%
Lis i zając	17	21	12	50	28.8%	14.7%	17.9%	18.6%
Słoń i lew	5	17	3	25	8.5%	11.9%	4.5%	9.3%
Overall	59	143	67	269	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



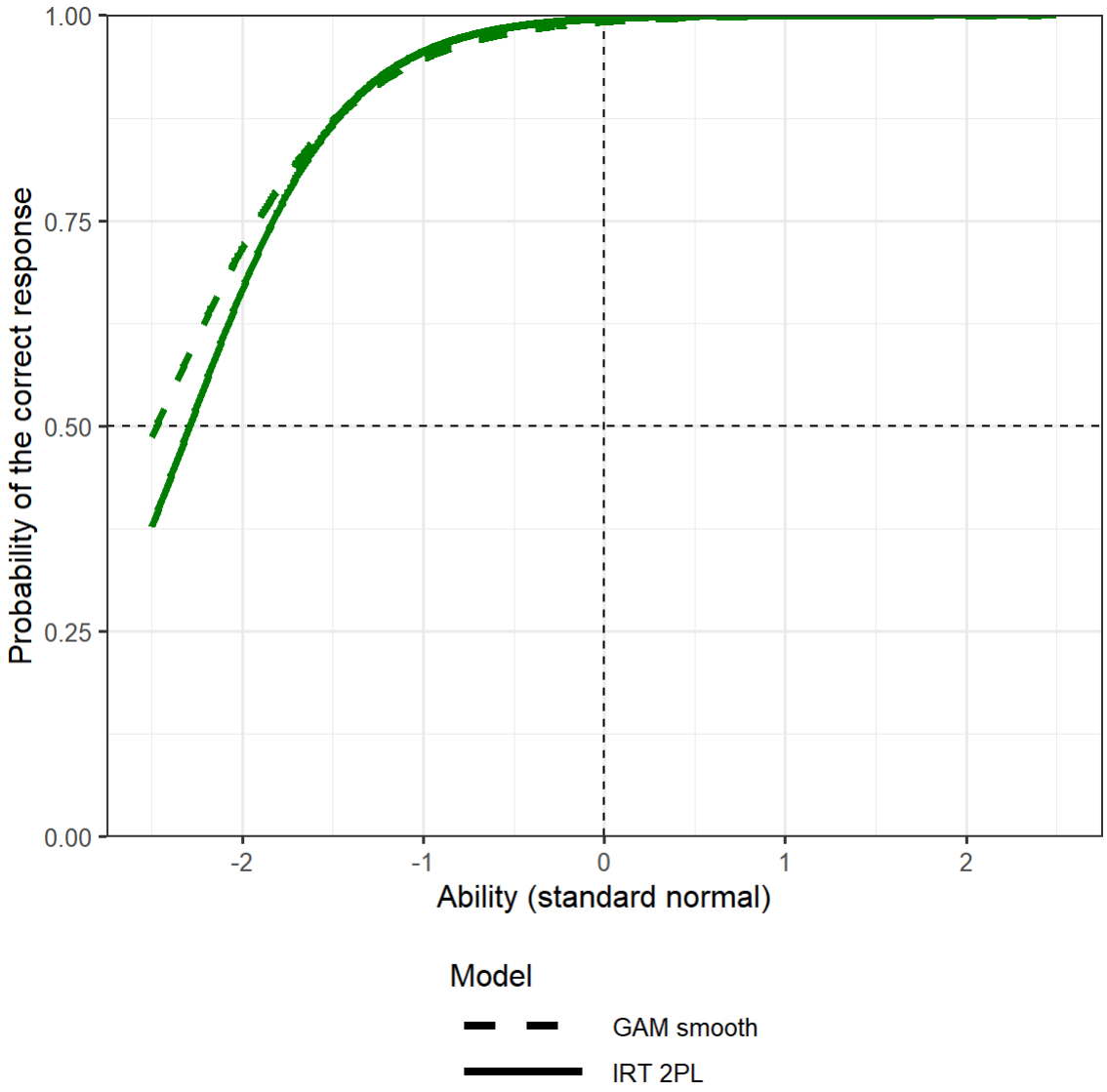
10.2.50.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	1.9	11.0	19.4	36.2	76.5	2.9	0.8
Middle quartiles	2.2	15.3	22.4	35.4	107.0	3.1	0.6
Upper quartile	8.6	17.9	26.3	36.4	132.3	3.3	0.5
Overall	1.9	15.2	22.8	35.6	132.3	3.1	0.7

10.2.51 S100010 (TIMSS)

10.2.51.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	96.64%	299	LLM simple setup	MC	0.29	0.20	2.38	−2.29	0.038

10.2.51.2 Gender DIF:

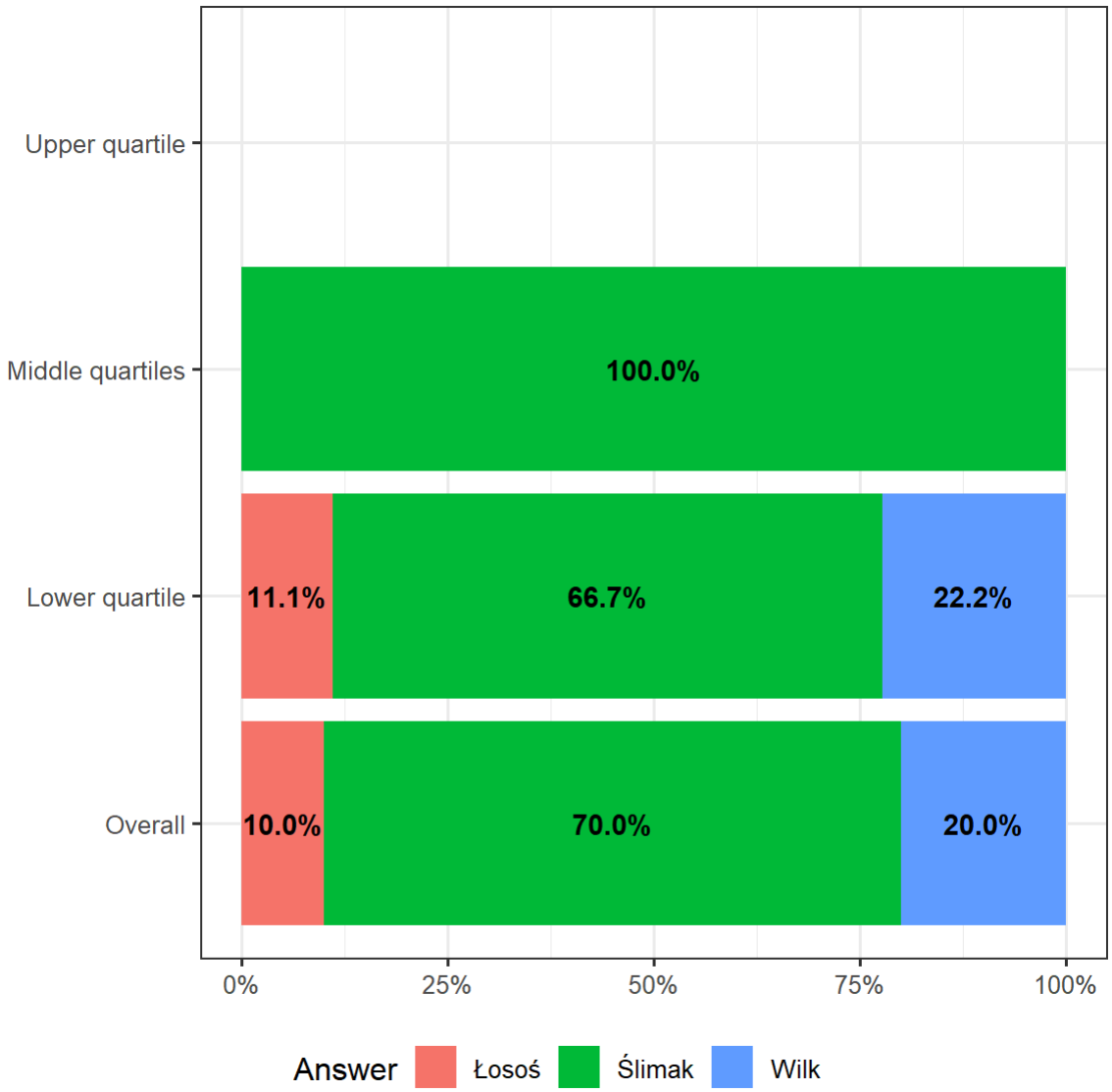
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.11	0.211	1.000	0.036	N/A

10.2.51.3 Frequency of answers:

Answer	Frequency	Percent
Sosna (correct)	288	96.6%
Wilk	2	0.7%
Łosoś	1	0.3%
Ślimak	7	2.3%
Overall	298	100.0%

10.2.51.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Wilk	2	0	0	2	22.2%	0.0%	0.0%	20.0%
Łosoś	1	0	0	1	11.1%	0.0%	0.0%	10.0%
Ślimak	6	1	0	7	66.7%	100.0%	0.0%	70.0%
Overall	9	1	0	10	100.0%	100.0%	0.0%	100.0%

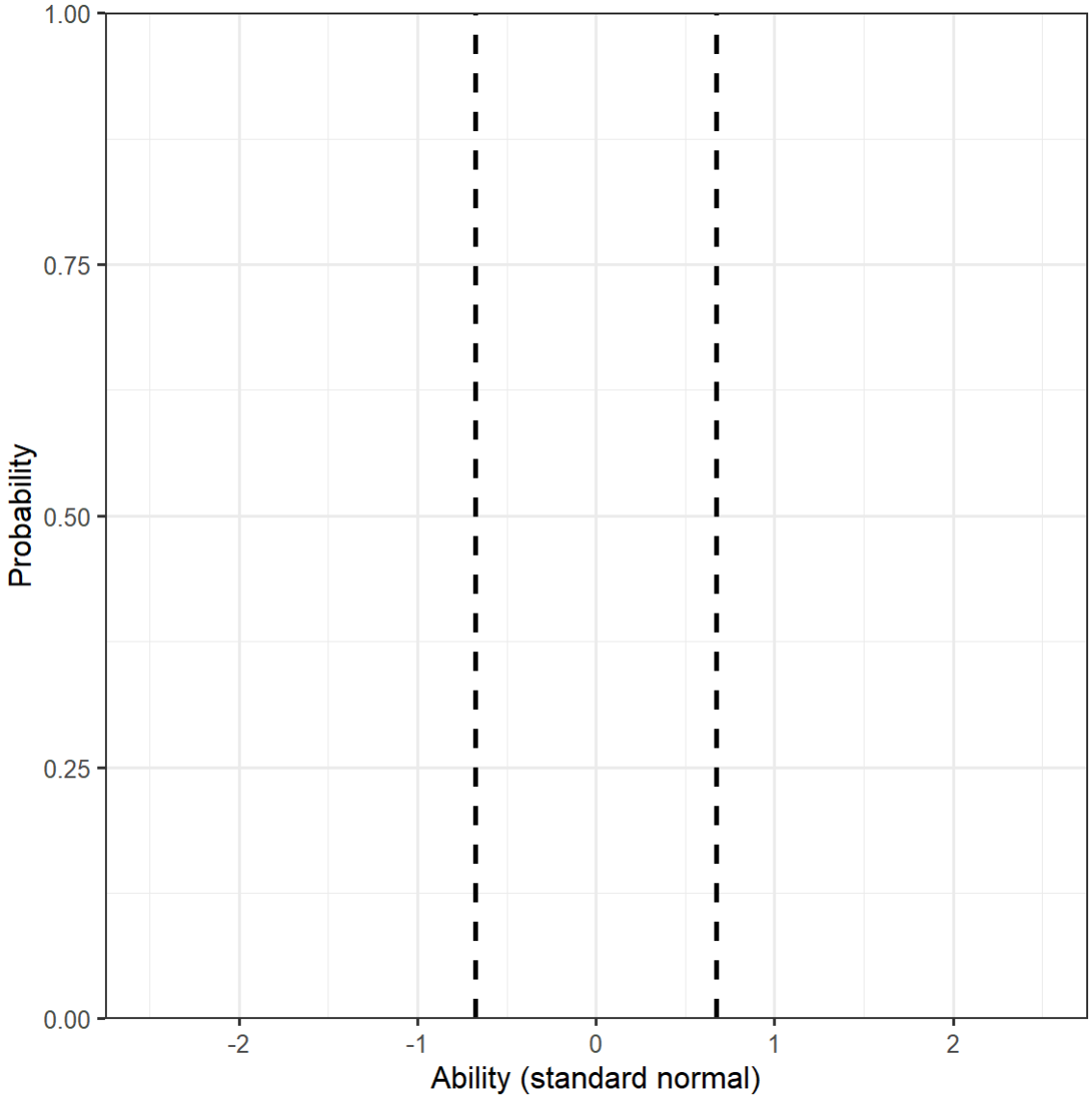



```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

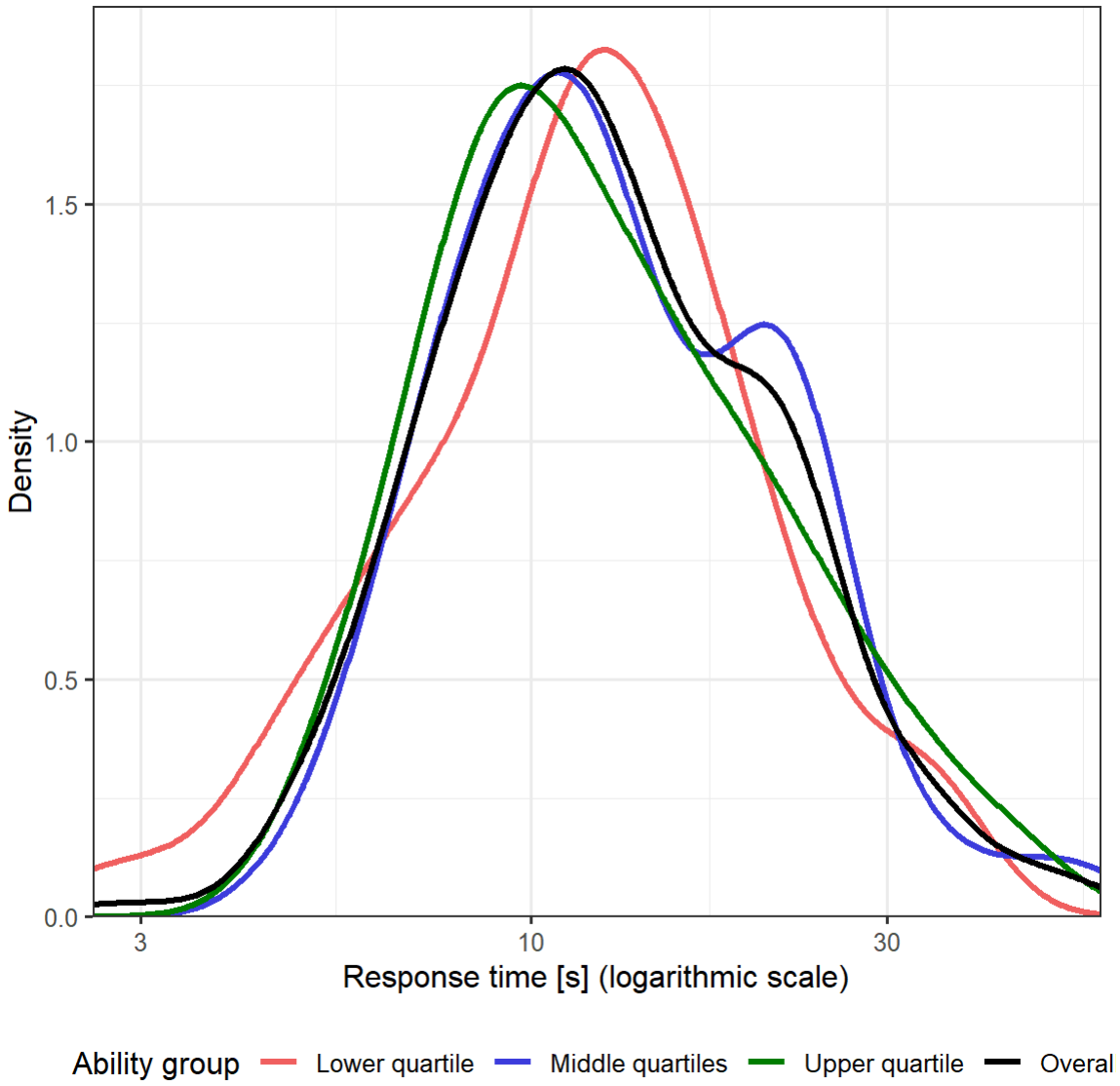
```
## Warning: Failed to fit group 1.
## Caused by error in `smooth.construct.cr.smooth.spec()`:
## ! x has insufficient unique values to support 10 knots: reduce k.
```

```
## Warning: Failed to fit group 2.
## Caused by error in `smooth.construct.cr.smooth.spec()`:
## ! x has insufficient unique values to support 10 knots: reduce k.
```

```
## Warning: Failed to fit group 3.
## Caused by error in `smooth.construct.cr.smooth.spec()`:
## ! x has insufficient unique values to support 10 knots: reduce k.
```



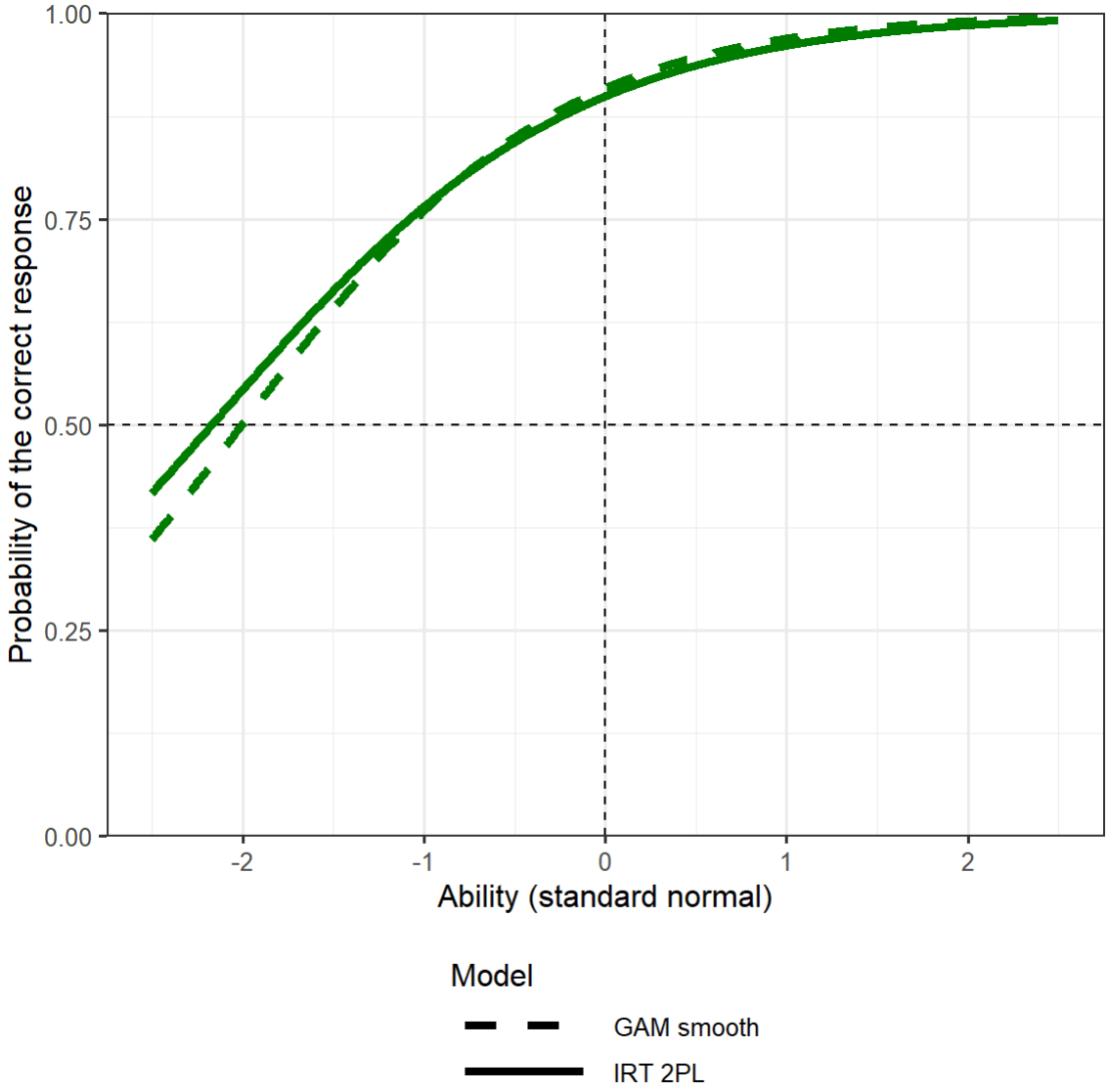
10.2.51.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.6	8.3	12.1	15.9	35.0	2.5	0.5
Middle quartiles	4.7	9.2	12.0	19.4	57.8	2.6	0.5
Upper quartile	5.1	8.7	11.5	17.9	45.6	2.5	0.5
Overall	2.6	8.8	12.0	18.5	57.8	2.5	0.5

10.2.52 S041010 (TIMSS)

10.2.52.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	85.90%	306	TIMSS	MC	0.35	0.32	1.01	−2.17	0.012

10.2.52.2 Gender DIF:

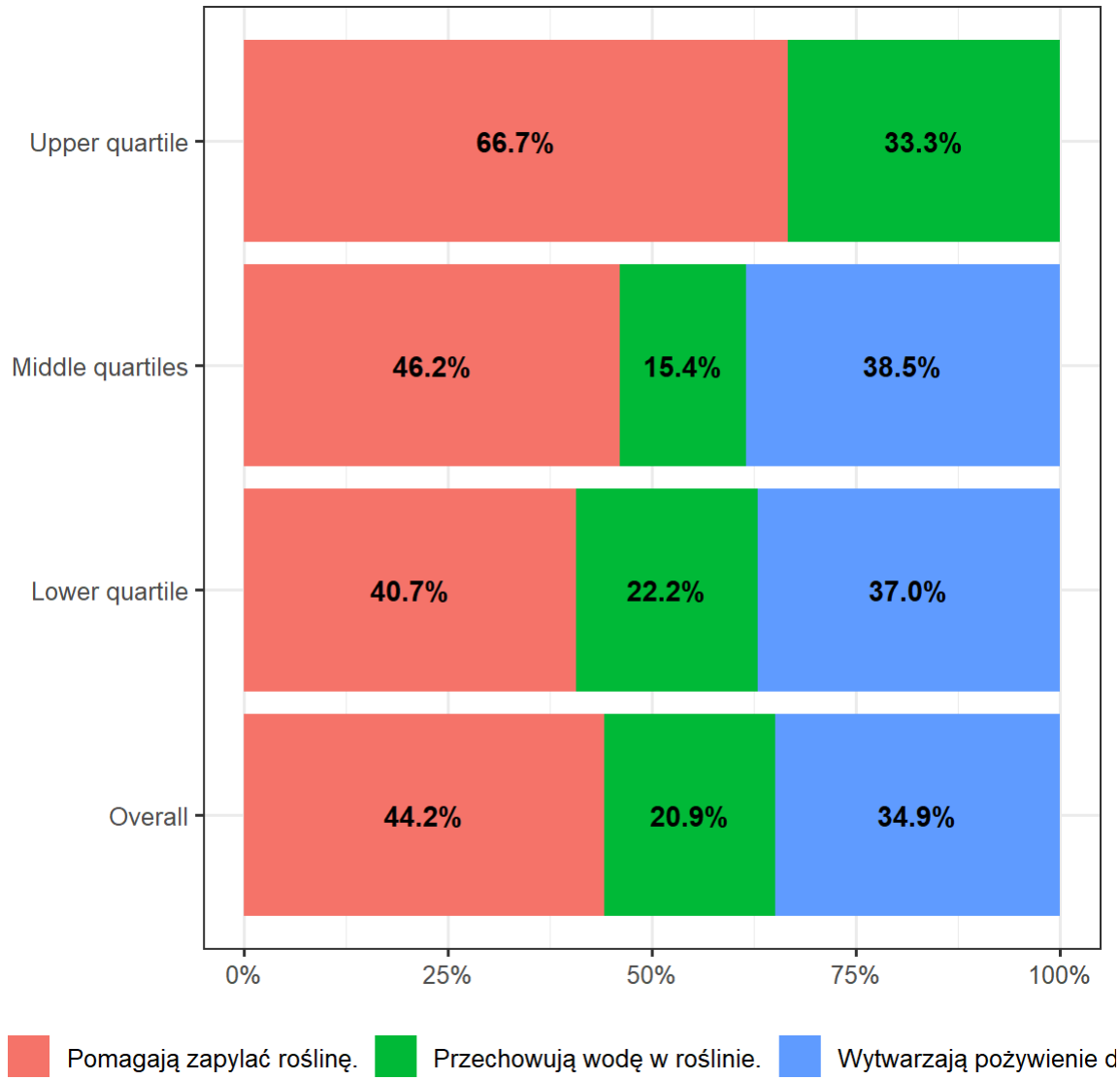
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
10.43	0.005	0.477	0.045	N/A

10.2.52.3 Frequency of answers:

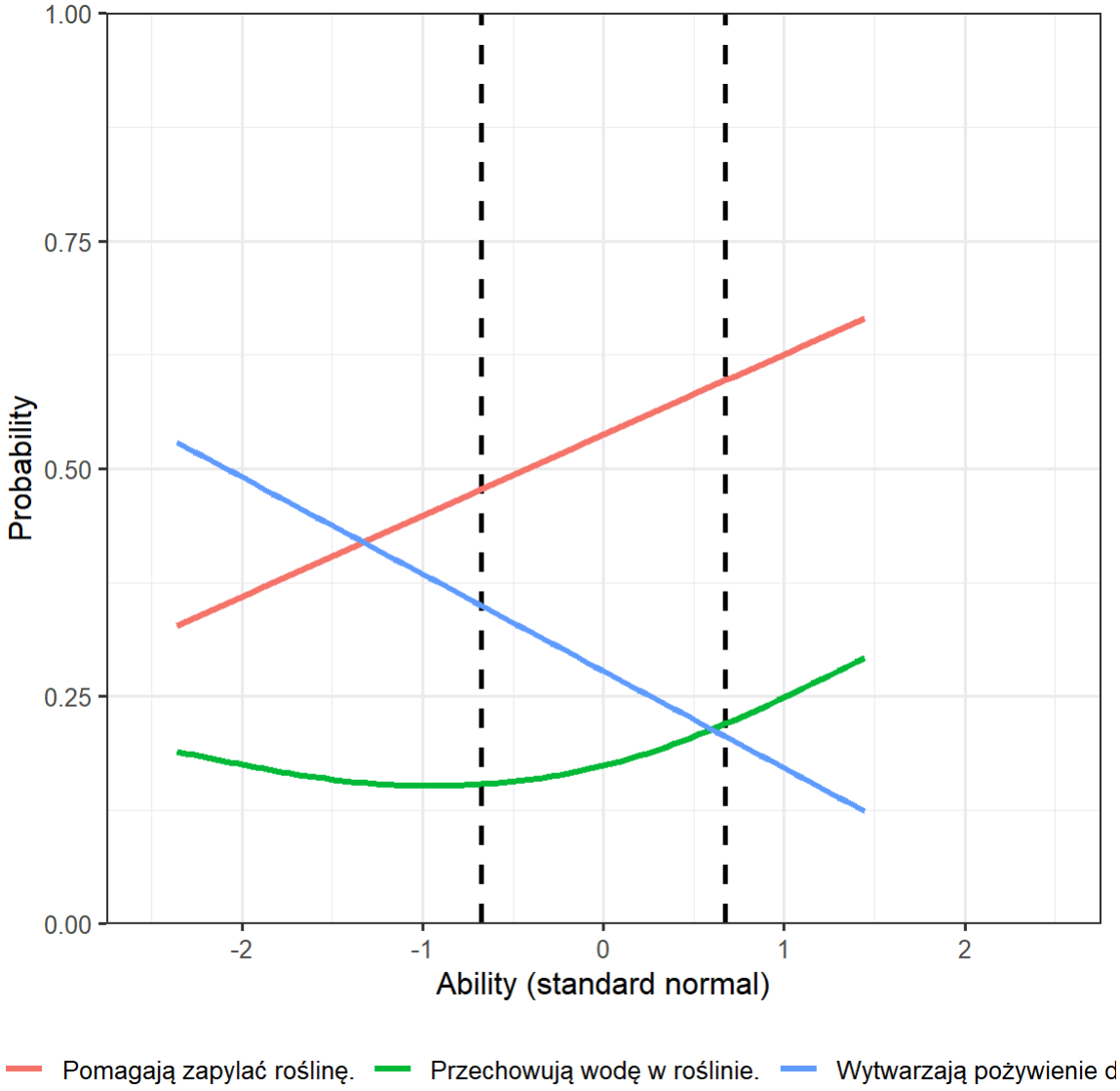
Answer	Frequency	Percent
Pomagają zapylać roślinę.	19	6.2%
Przechowują wodę w roślinie.	9	3.0%
Wytwarzają nowe rośliny. (correct)	262	85.9%
Wytwarzają pożywienie dla rośliny.	15	4.9%
Overall	305	100.0%

10.2.52.4 Distractors functioning:

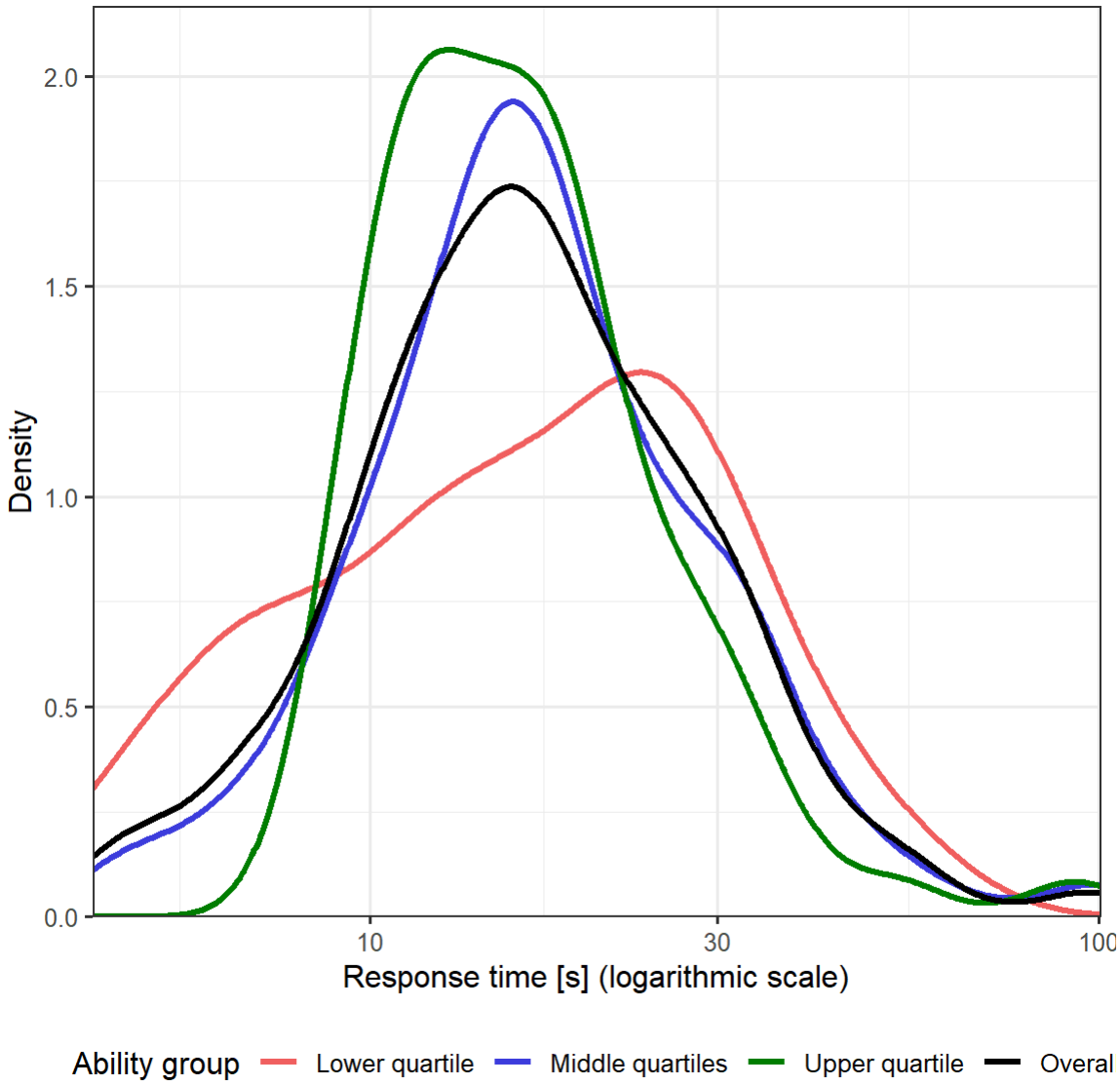
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Pomagają zapylać roślinę.	11	6	2	19	40.7%	46.2%	66.7%	44.2%
Przechowują wodę w roślinie.	6	2	1	9	22.2%	15.4%	33.3%	20.9%
Wytwarzają pożywienie dla rośliny.	10	5	0	15	37.0%	38.5%	0.0%	34.9%
Overall	27	13	3	43	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



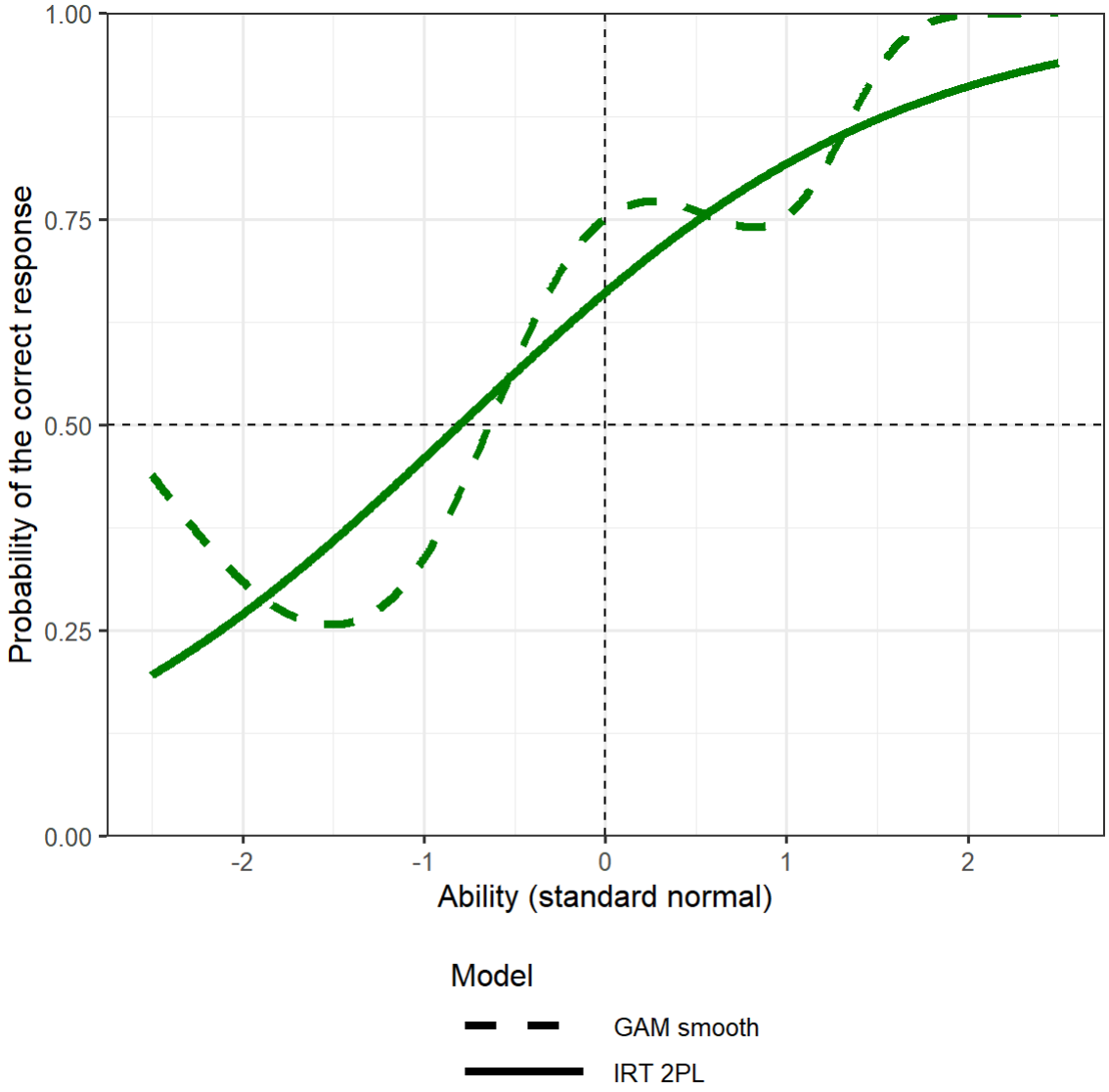
10.2.52.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.2	9.9	16.3	25.2	56.9	2.8	0.7
Middle quartiles	4.5	12.2	16.4	23.4	100.5	2.8	0.5
Upper quartile	8.4	11.5	15.5	20.0	93.0	2.8	0.4
Overall	4.2	11.5	16.1	23.7	100.5	2.8	0.6

10.2.53 S041209 (LLM)

10.2.53.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	64.77%	299	TIMSS	MC	0.38	0.30	0.83	−0.80	0.012

10.2.53.2 Gender DIF:

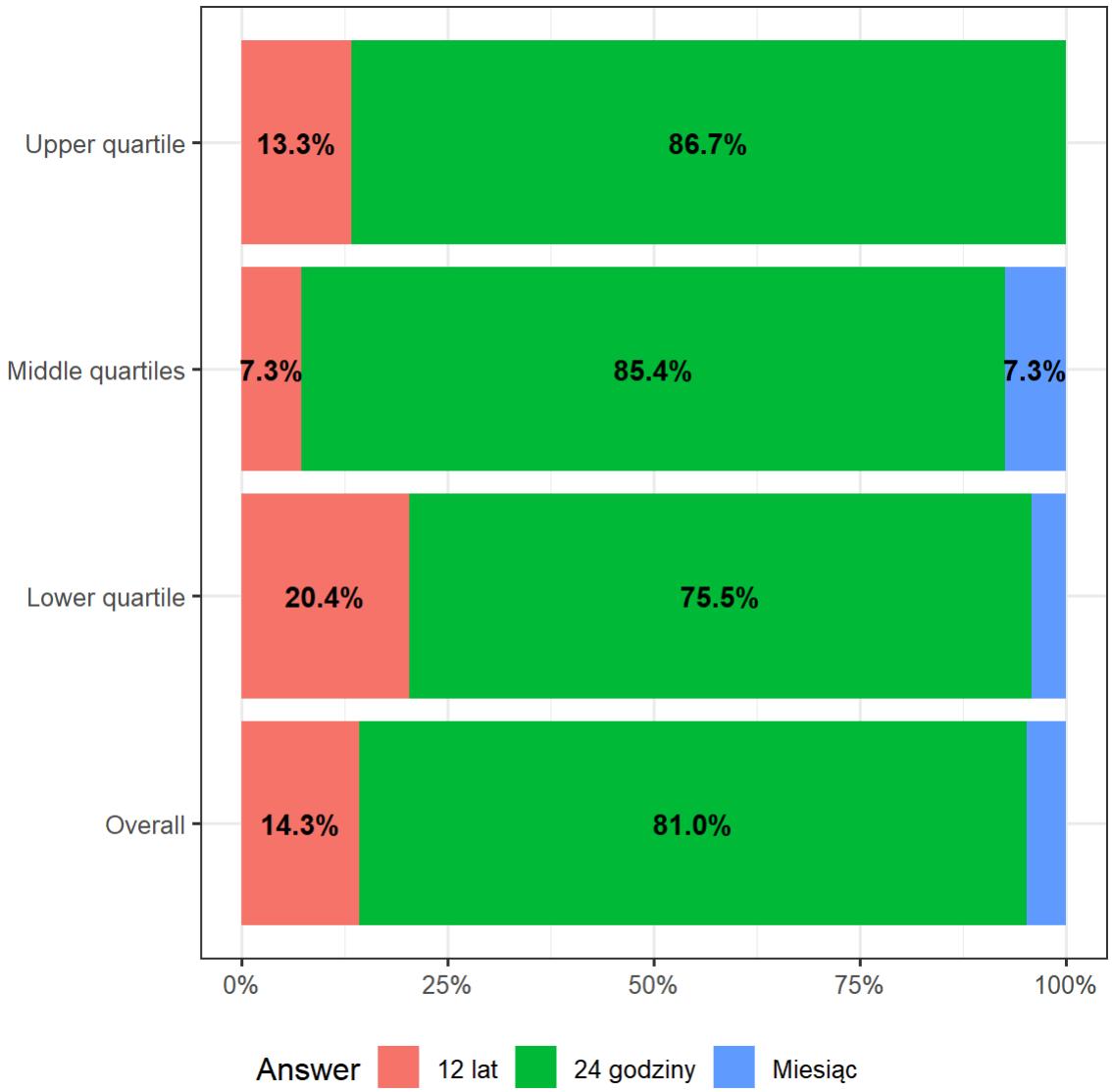
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.21	0.898	1.000	0.001	N/A

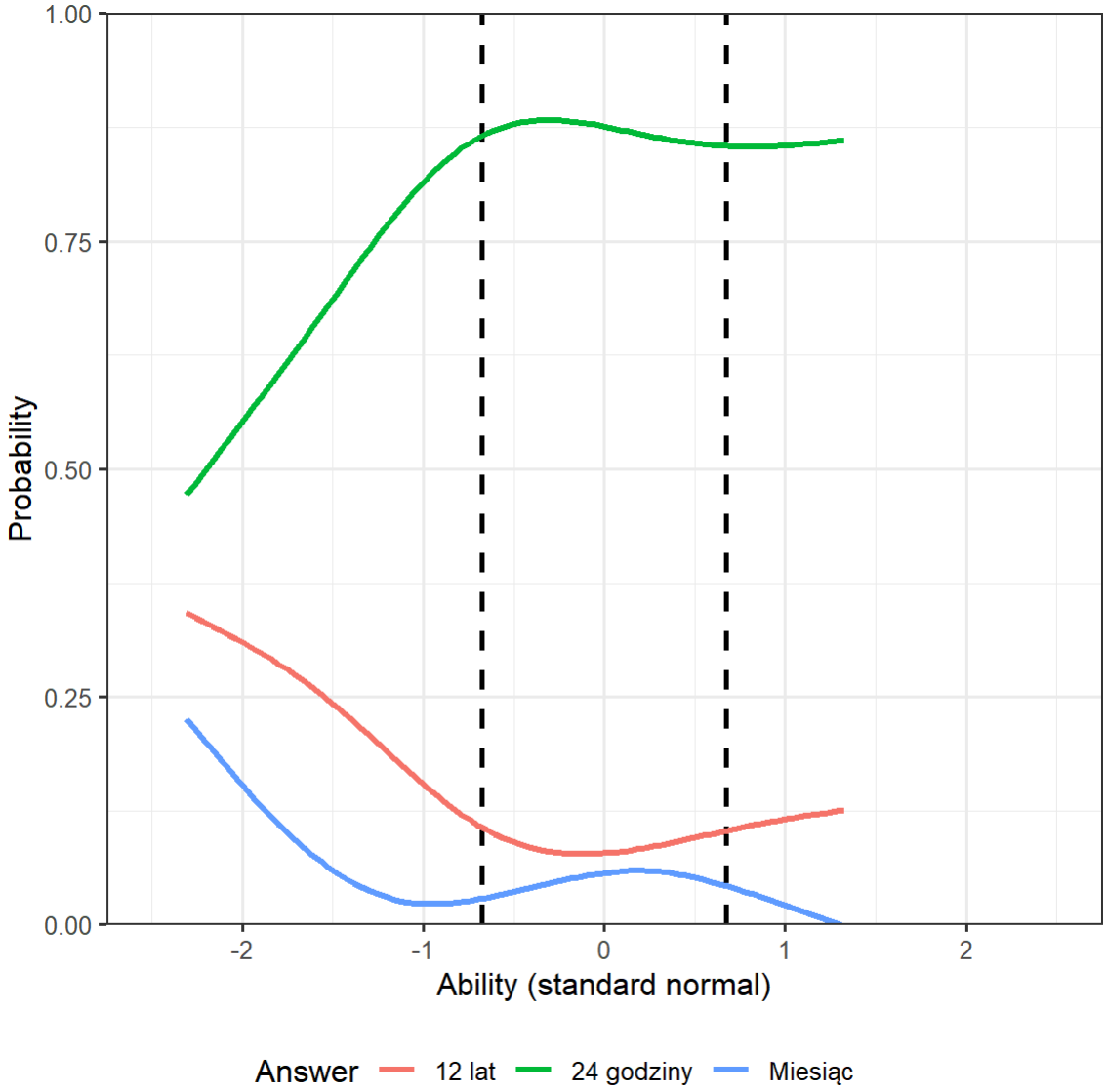
10.2.53.3 Frequency of answers:

Answer	Frequency	Percent
12 lat	15	5.0%
24 godziny	85	28.5%
Miesiąc	5	1.7%
Rok (correct)	193	64.8%
Overall	298	100.0%

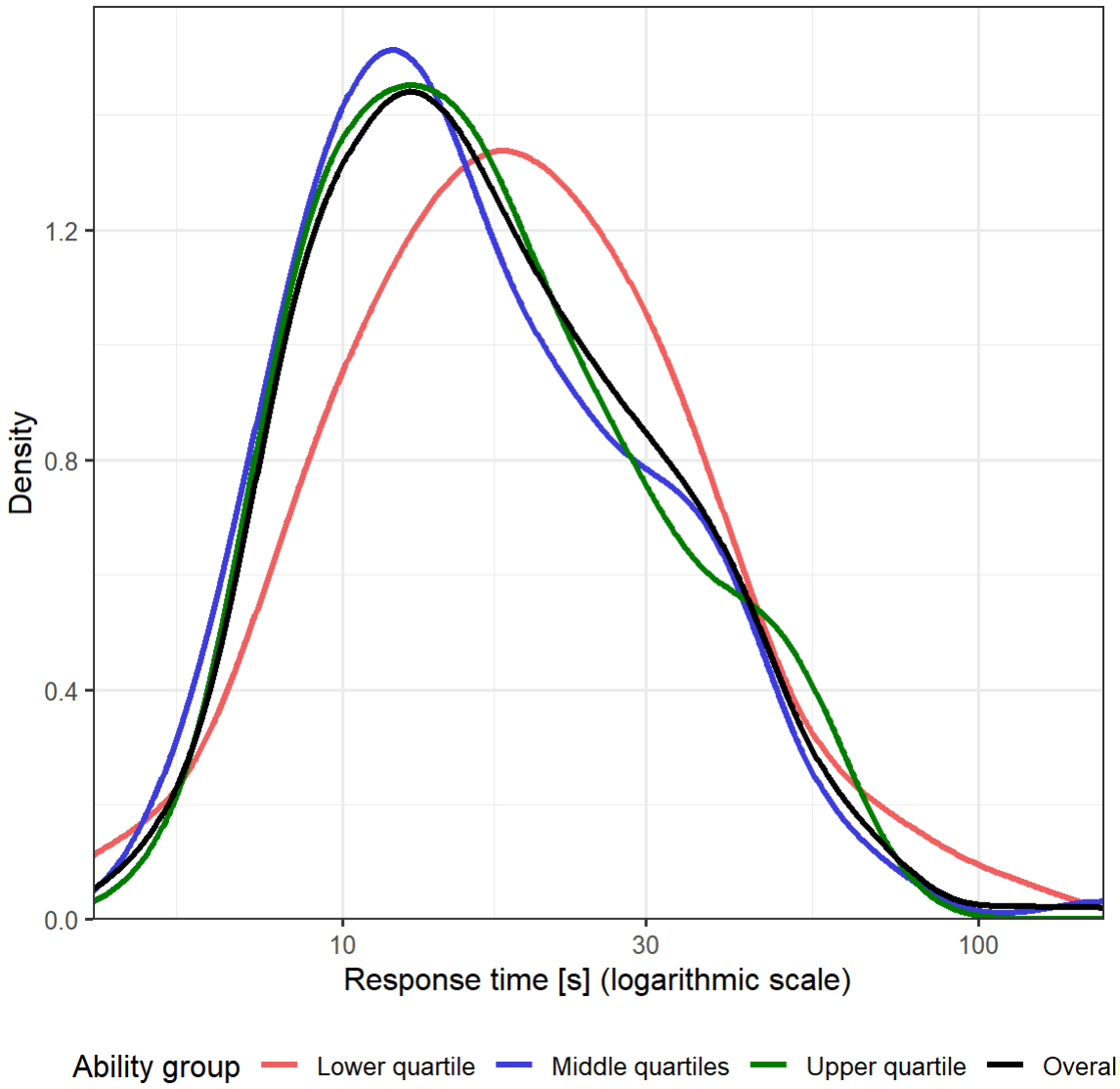
10.2.53.4 Distractors functioning:

	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
12 lat	10	3	2	15	20.4%	7.3%	13.3%	14.3%
24 godziny	37	35	13	85	75.5%	85.4%	86.7%	81.0%
Miesiąc	2	3	0	5	4.1%	7.3%	0.0%	4.8%
Overall	49	41	15	105	100.0%	100.0%	100.0%	100.0%


Upper quartile
Middle quartiles
Lower quartile
Overall
13.3%
86.7%
7.3%
85.4%
7.3%
20.4%
75.5%
14.3%
81.0%
0%
25%
50%
75%
100%
Answer
12 lat
24 godziny
Miesiąc



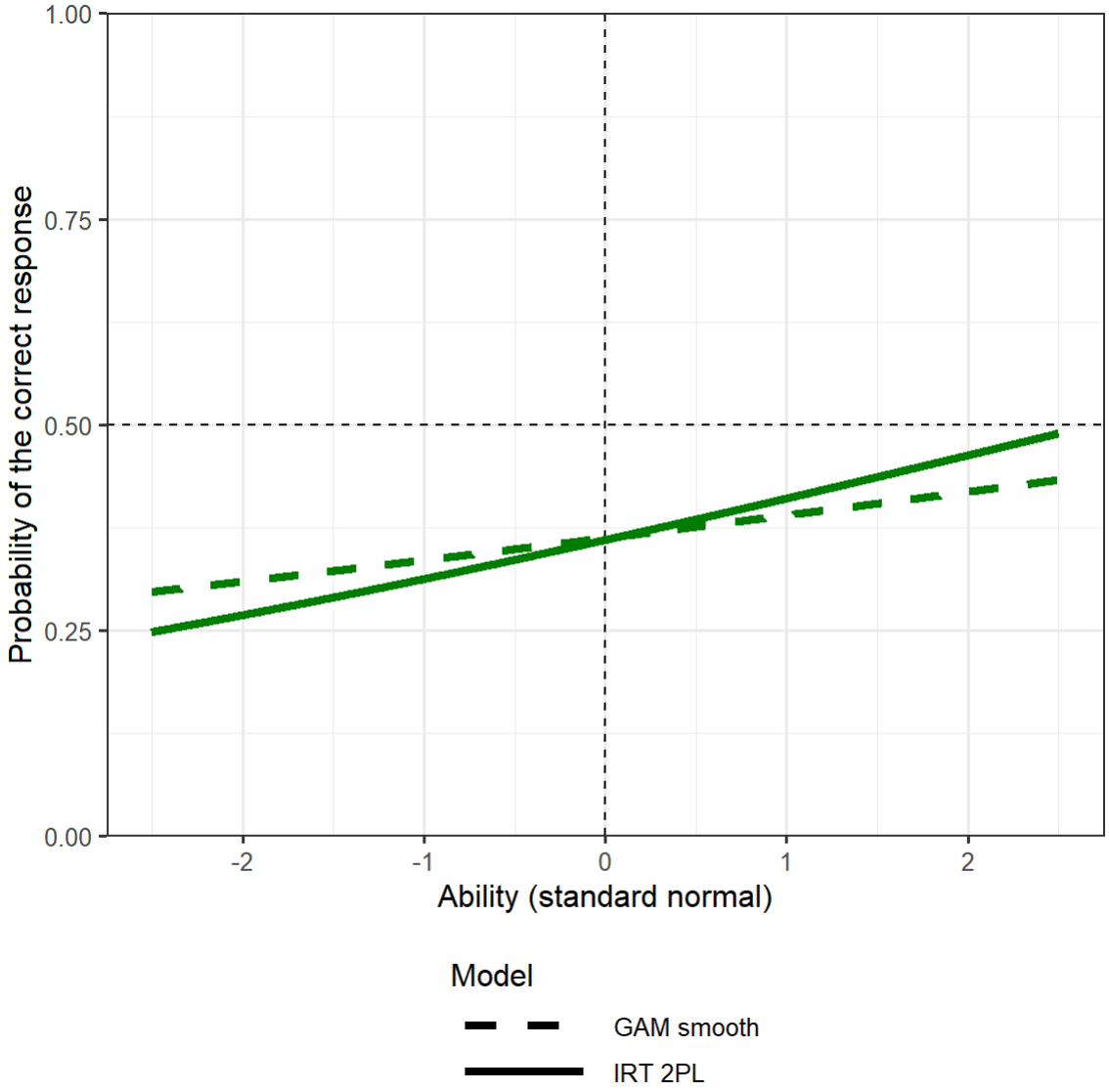
10.2.53.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.0	11.8	18.4	28.0	111.4	2.9	0.6
Middle quartiles	5.1	10.0	14.1	24.7	157.2	2.8	0.6
Upper quartile	5.3	10.5	15.9	24.5	58.9	2.8	0.6
Overall	4.0	10.7	15.8	25.3	157.2	2.8	0.6

10.2.54 S090048 (TIMSS)

10.2.54.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	36.24%	299	LLM teachers' selection	MC	0.05	0.03	0.21	2.70	0.0(

10.2.54.2 Gender DIF:

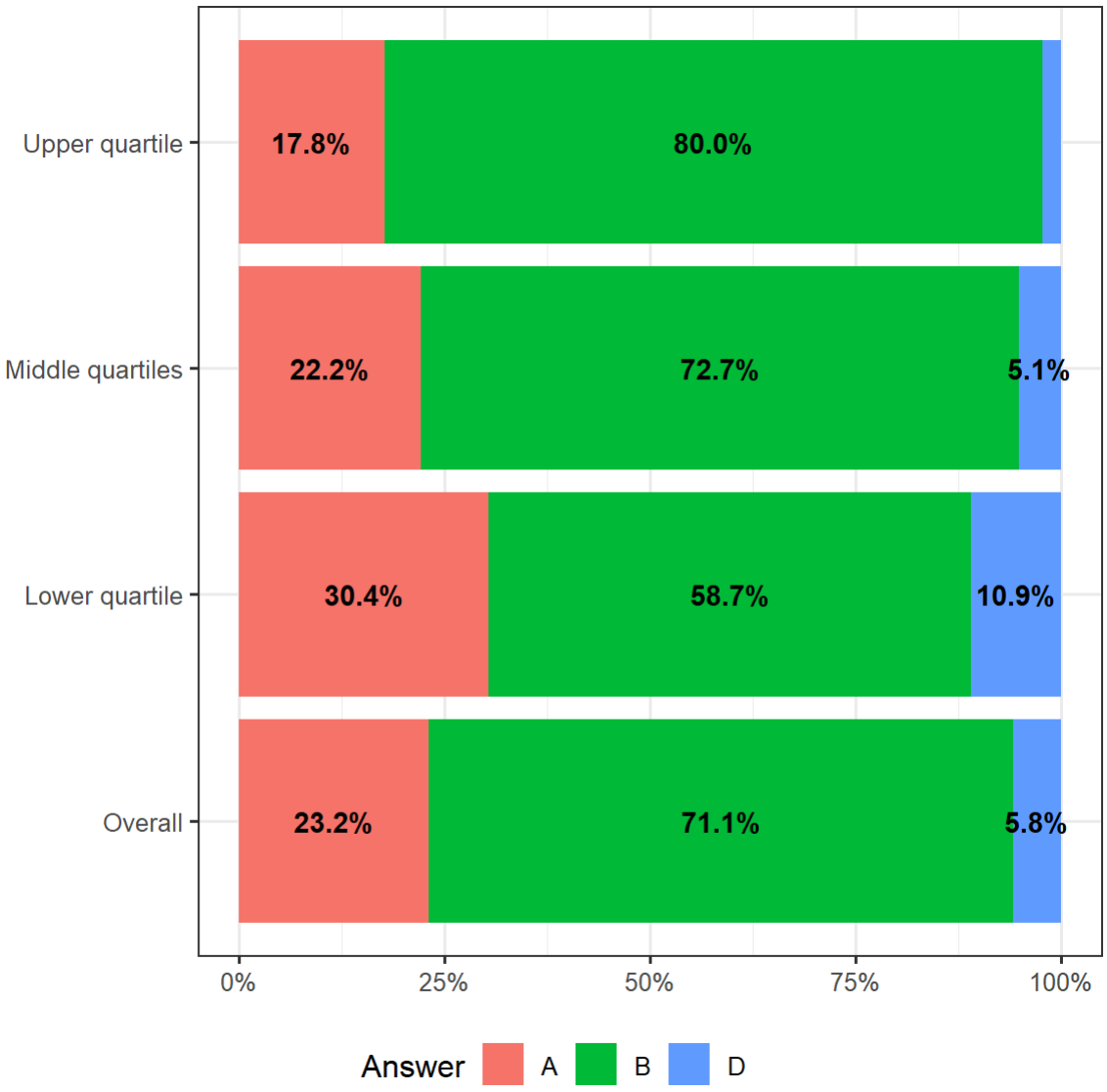
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.96	0.376	1.000	0.006	N/A

10.2.54.3 Frequency of answers:

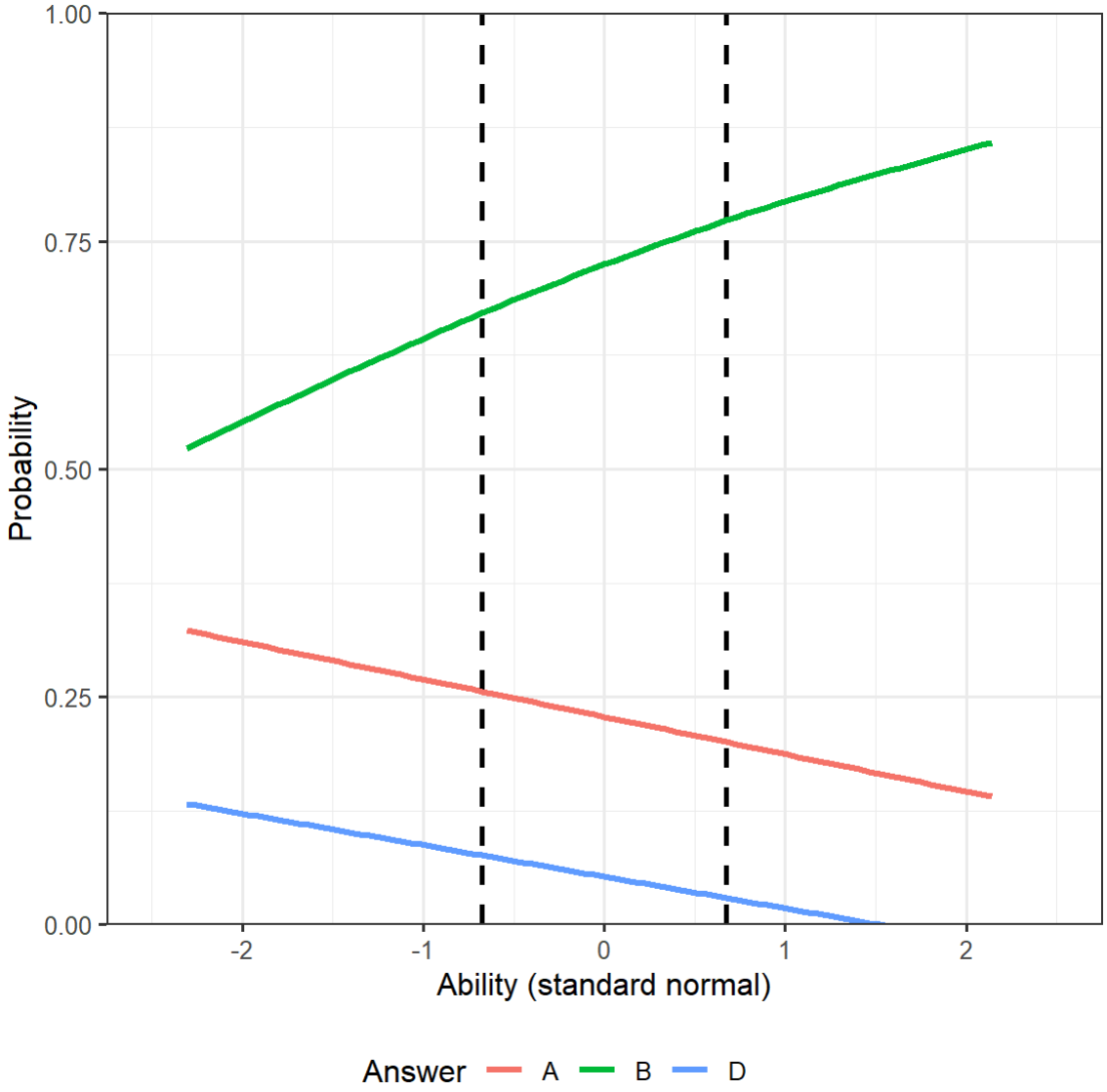
Answer	Frequency	Percent
A	44	14.8%
B	135	45.3%
C (correct)	108	36.2%
D	11	3.7%
Overall	298	100.0%

10.2.54.4 Distractors functioning:

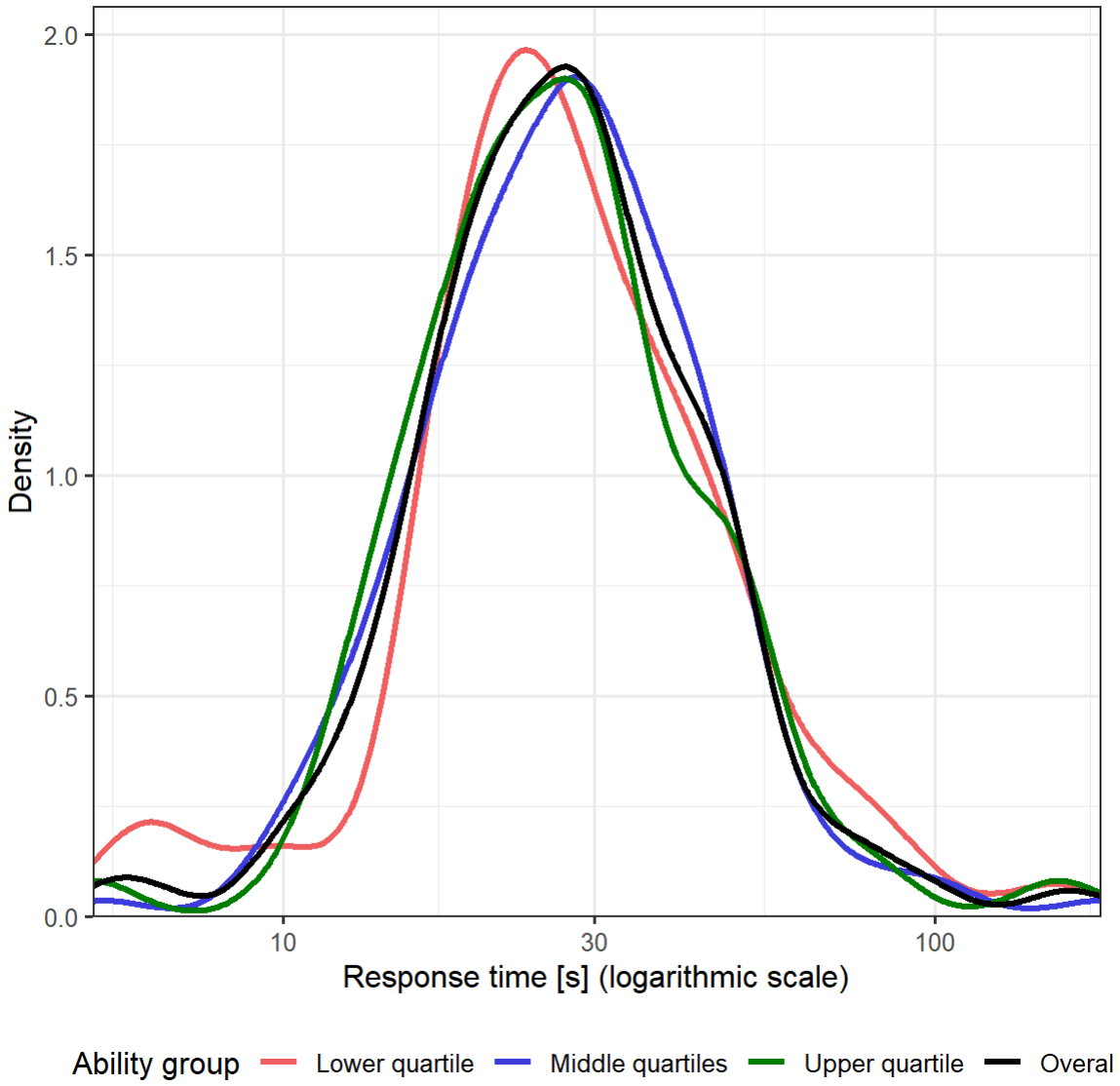
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
A	14	22	8	44	30.4%	22.2%	17.8%	23.2%
B	27	72	36	135	58.7%	72.7%	80.0%	71.1%
D	5	5	1	11	10.9%	5.1%	2.2%	5.8%
Overall	46	99	45	190	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



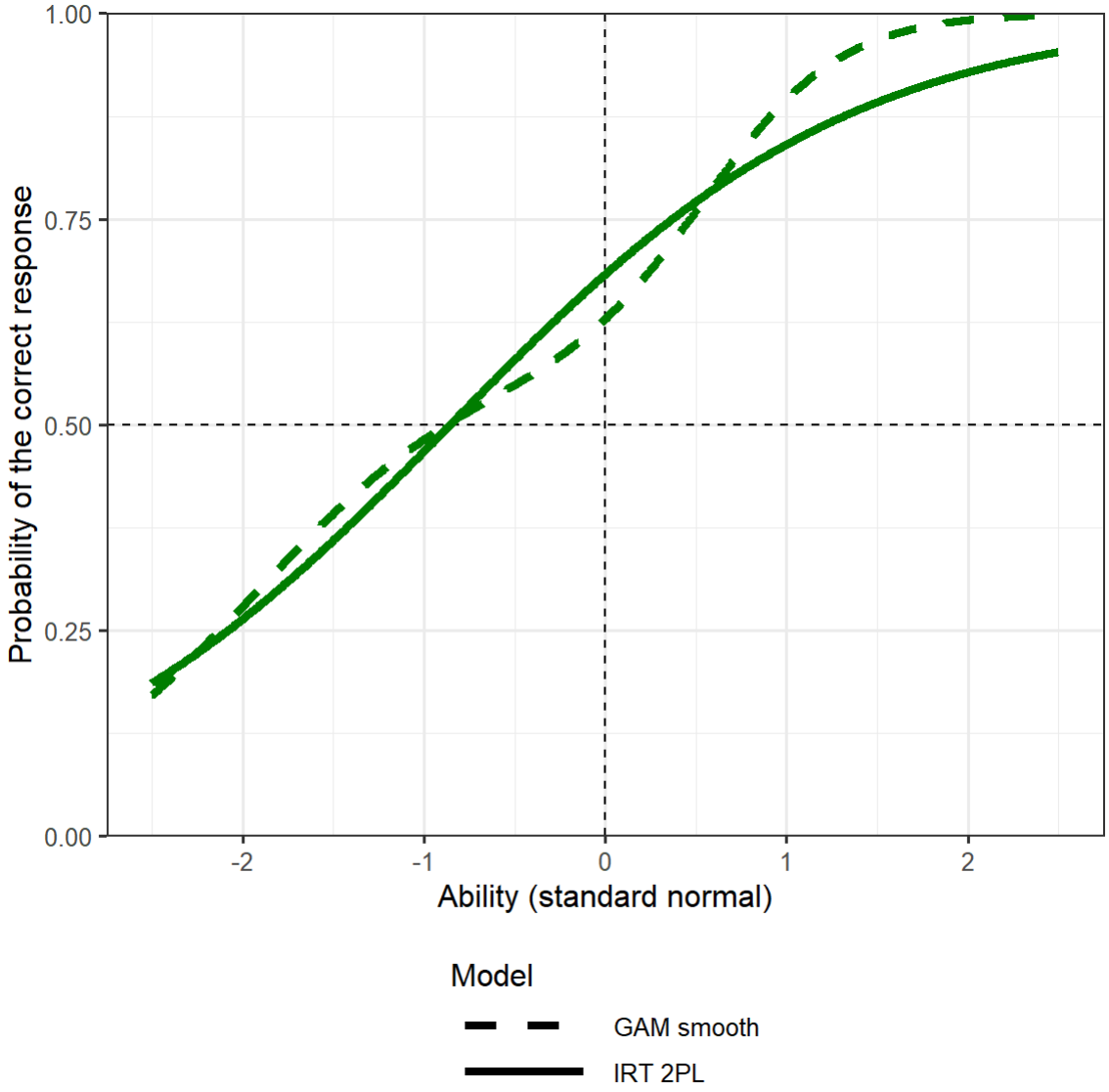
10.2.54.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	6.0	19.7	25.3	36.6	154.0	3.3	0.6
Middle quartiles	5.3	19.5	27.4	37.6	179.4	3.3	0.5
Upper quartile	5.1	19.3	26.2	33.8	155.2	3.3	0.5
Overall	5.1	19.5	26.6	37.0	179.4	3.3	0.5

10.2.55 S090053 (TIMSS)

10.2.55.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	65.46%	304	LLM teachers' selection	MC	0.38	0.25	0.90	−0.85	0.03

10.2.55.2 Gender DIF:

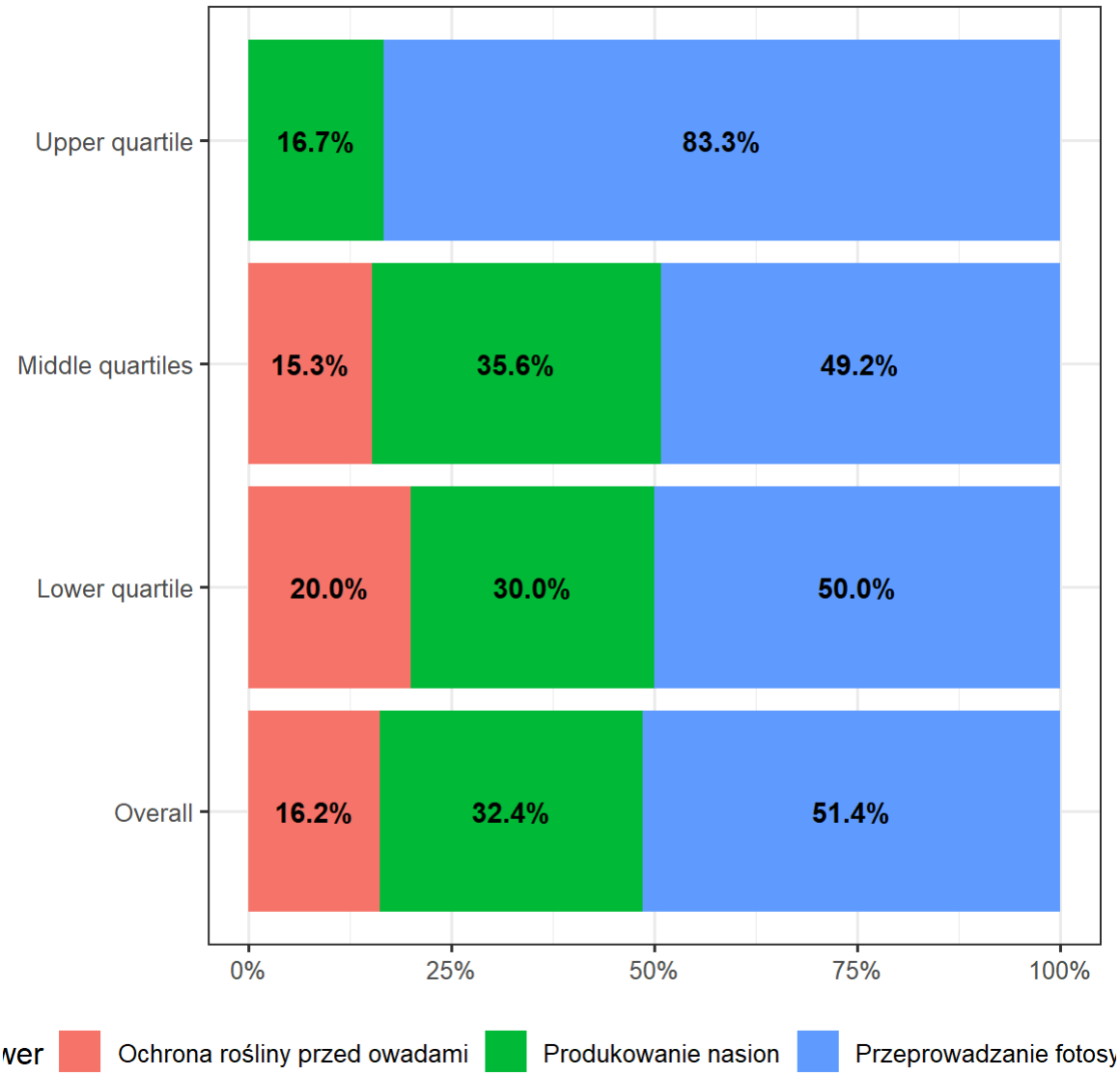
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.07	0.966	1.000	0.000	N/A

10.2.55.3 Frequency of answers:

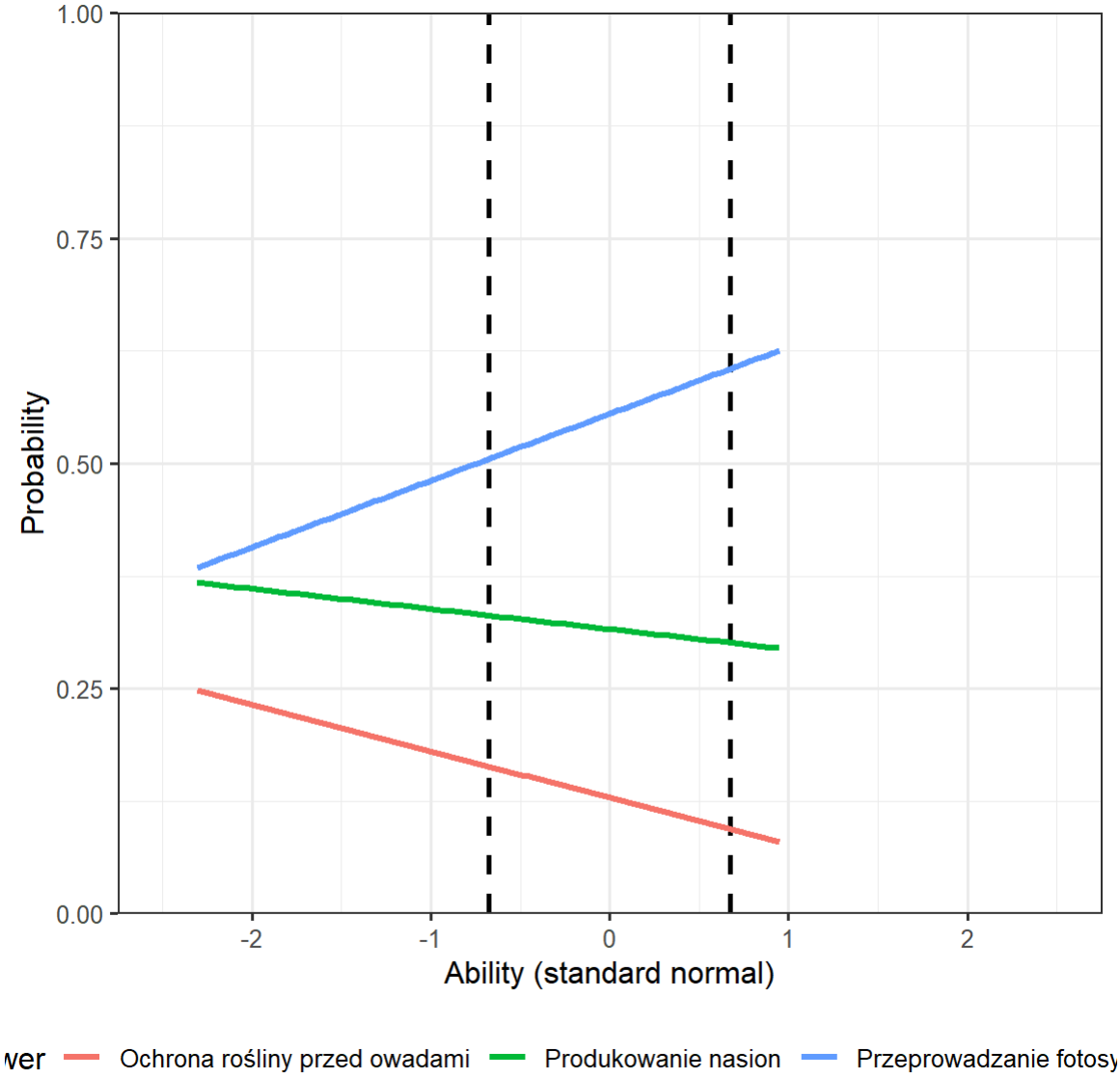
Answer	Frequency	Percent
Ochrona rośliny przed owadami	17	5.6%
Produkowanie nasion	34	11.2%
Przeprowadzanie fotosyntezy	54	17.8%
Wchłanianie wody i składników mineralnych z gleby (correct)	199	65.5%
Overall	304	100.0%

10.2.55.4 Distractors functioning:

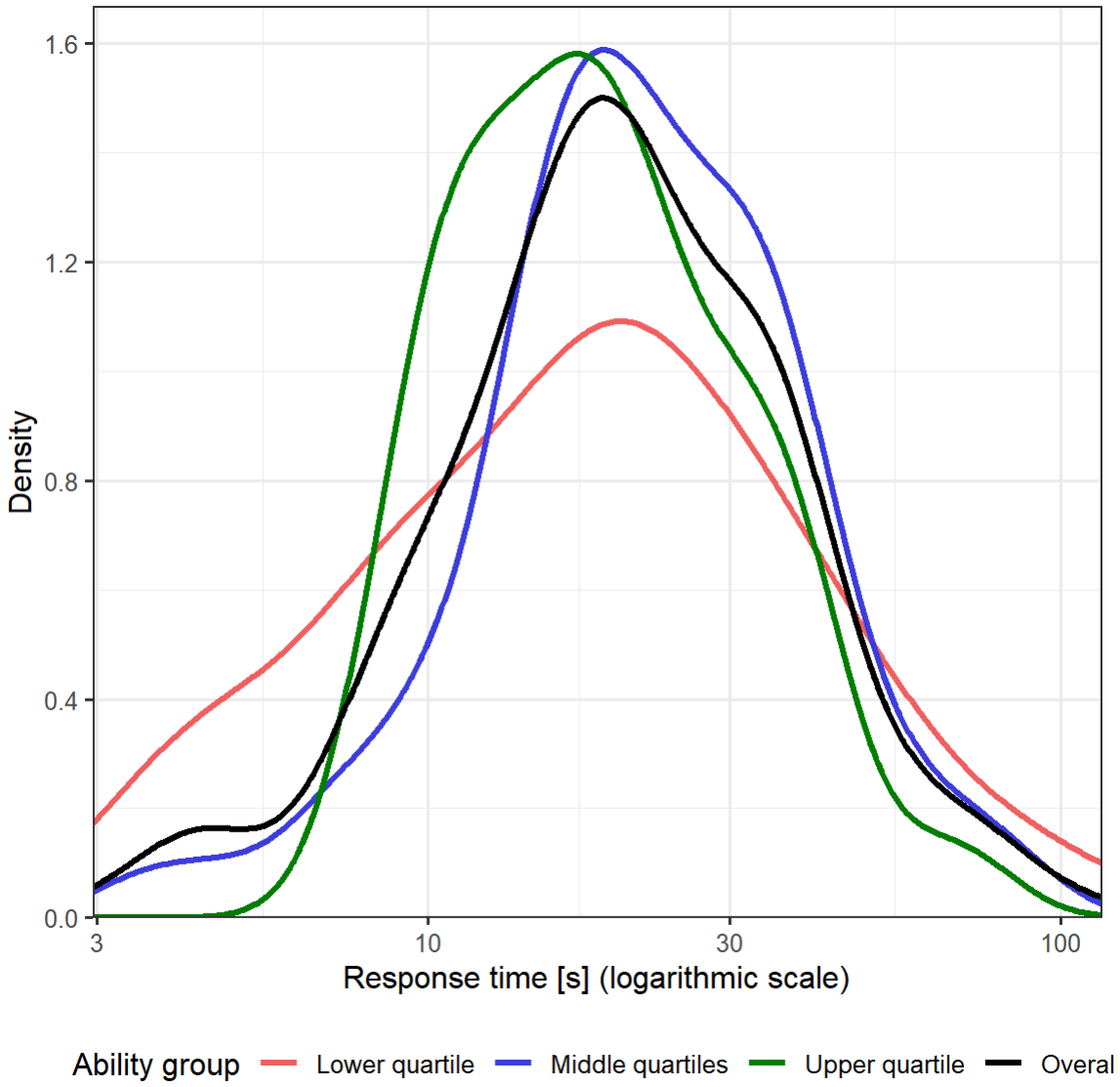
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Ochrona rośliny przed owadami	8	9	0	17	20.0%	15.3%	0.0%	16.2%
Produkowanie nasion	12	21	1	34	30.0%	35.6%	16.7%	32.4%
Przeprowadzanie fotosyntezy	20	29	5	54	50.0%	49.2%	83.3%	51.4%
Overall	40	59	6	105	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



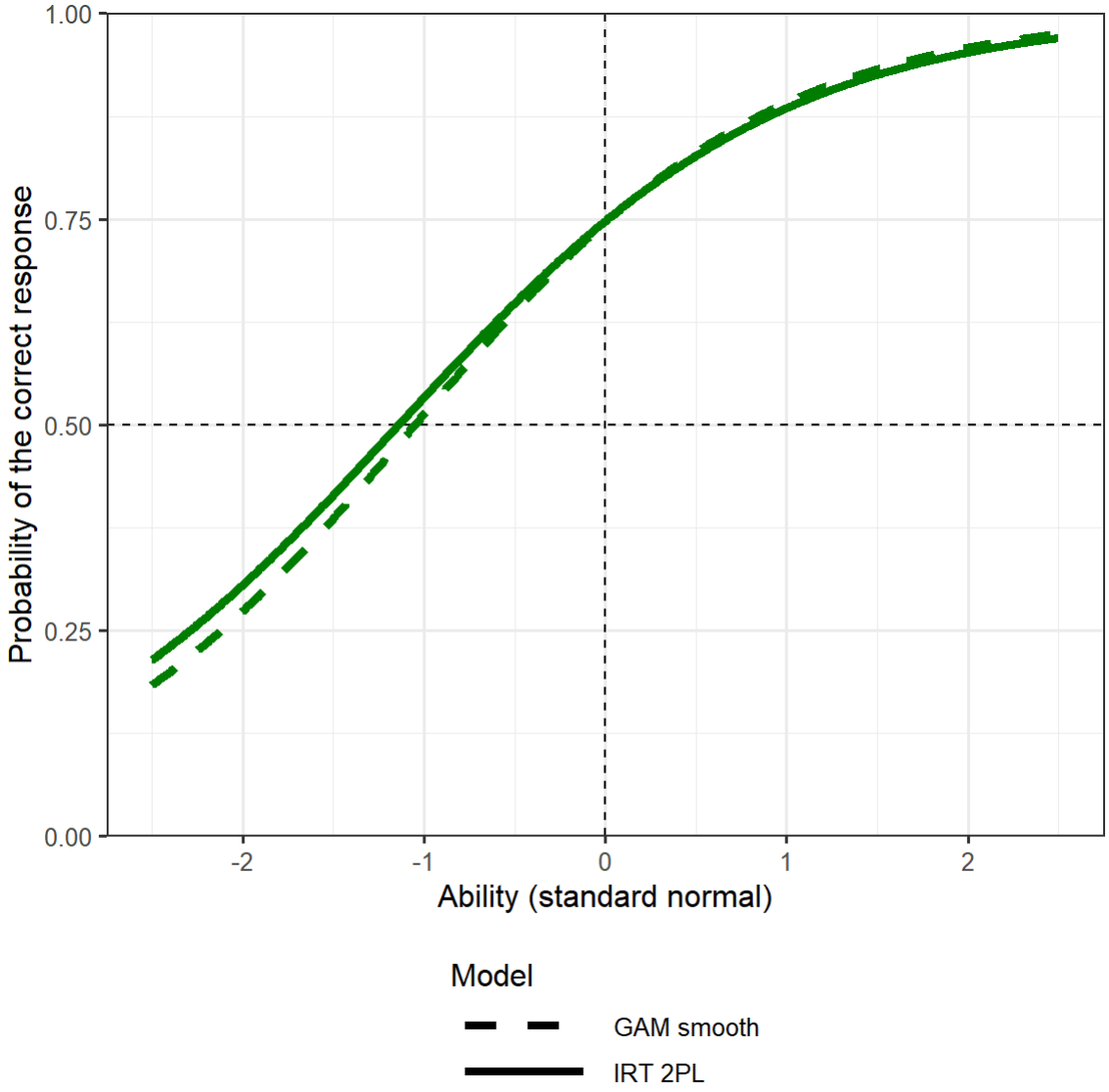
10.2.55.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	9.7	18.6	30.2	116.0	2.9	0.8
Middle quartiles	3.2	15.6	21.0	32.9	94.5	3.1	0.6
Upper quartile	7.9	12.4	17.8	26.9	72.6	2.9	0.5
Overall	3.0	13.4	19.7	31.1	116.0	3.0	0.6

10.2.56 S090018 (LLM)

10.2.56.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	71.05%	304	LLM teachers' selection	MC	0.37	0.26	0.95	−1.14	0.02

10.2.56.2 Gender DIF:

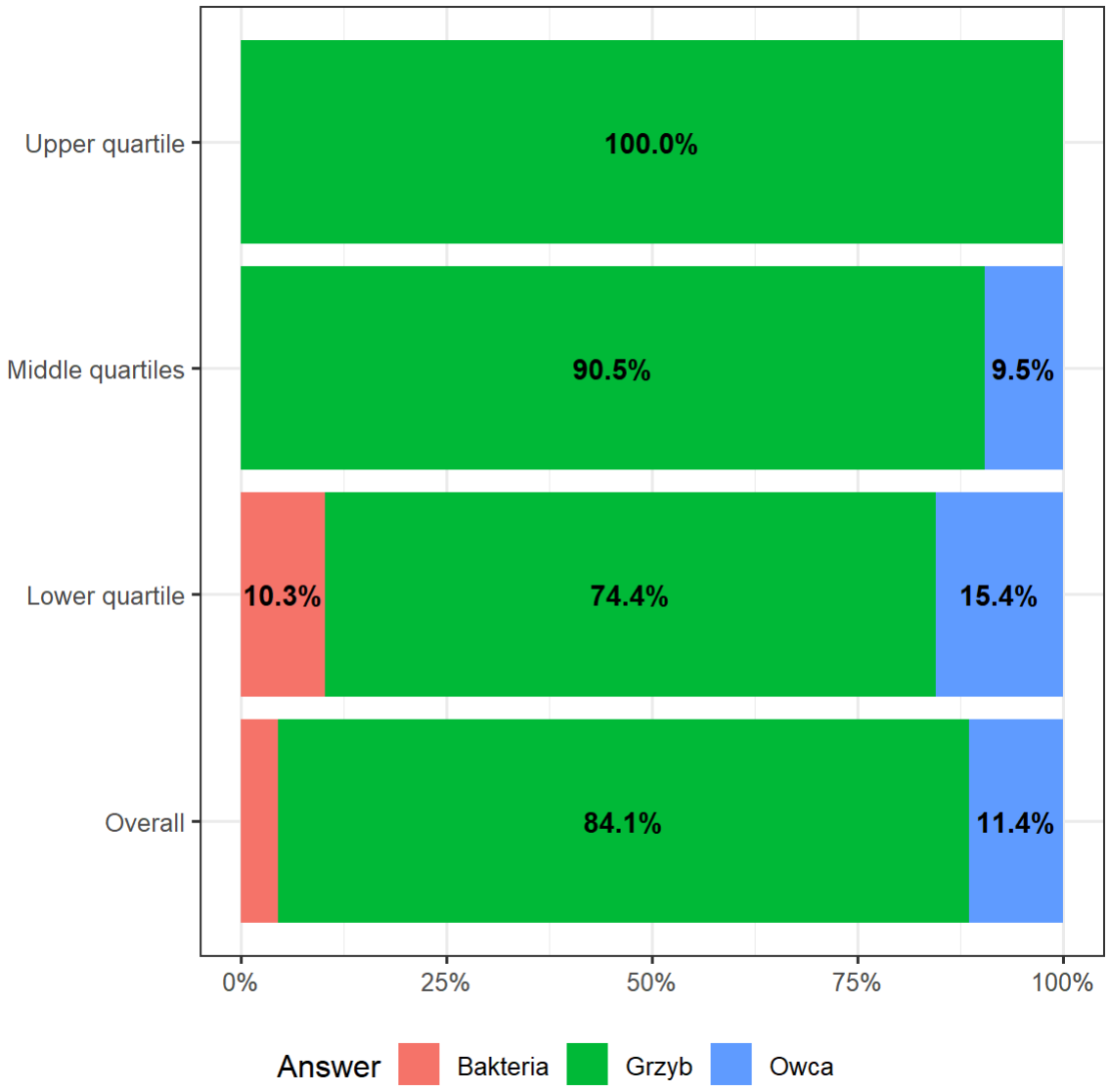
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
9.45	0.009	0.754	0.029	N/A

10.2.56.3 Frequency of answers:

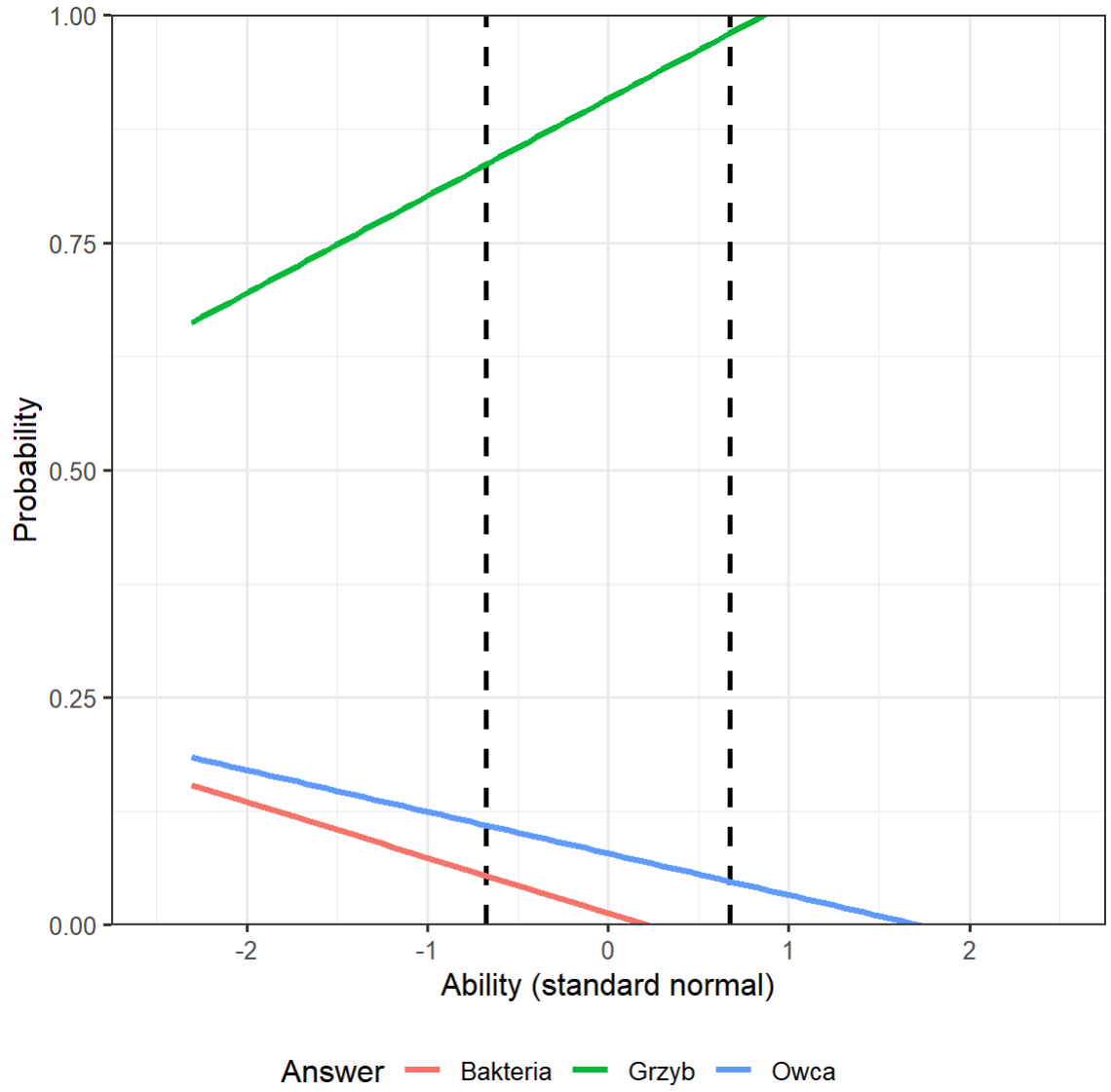
Answer	Frequency	Percent
Bakteria	4	1.3%
Grzyb	74	24.3%
Owca	10	3.3%
Sosna (correct)	216	71.1%
Overall	304	100.0%

10.2.56.4 Distractors functioning:

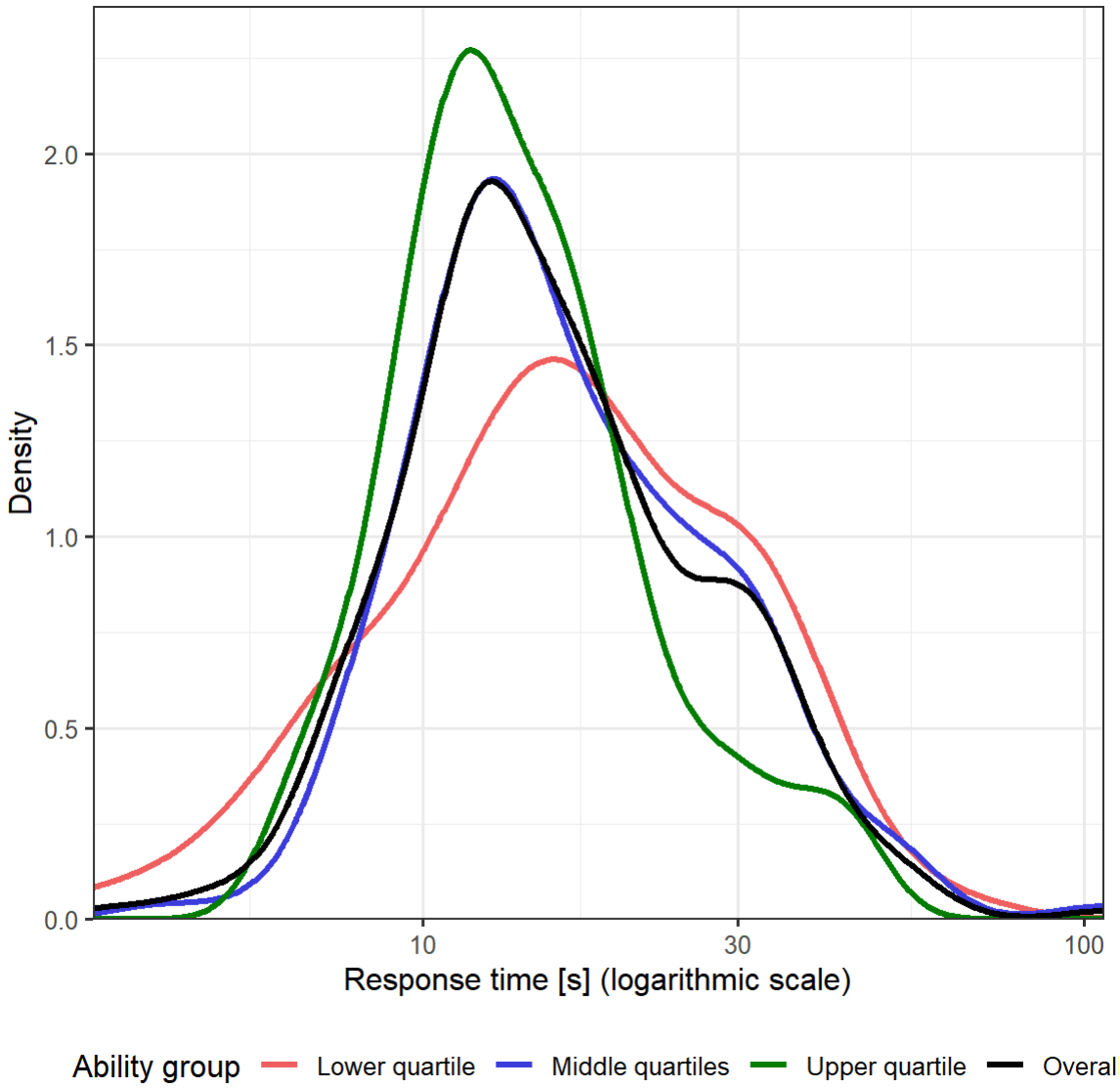
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Bakteria	4	0	0	4	10.3%	0.0%	0.0%	4.5%
Grzyb	29	38	7	74	74.4%	90.5%	100.0%	84.1%
Owca	6	4	0	10	15.4%	9.5%	0.0%	11.4%
Overall	39	42	7	88	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



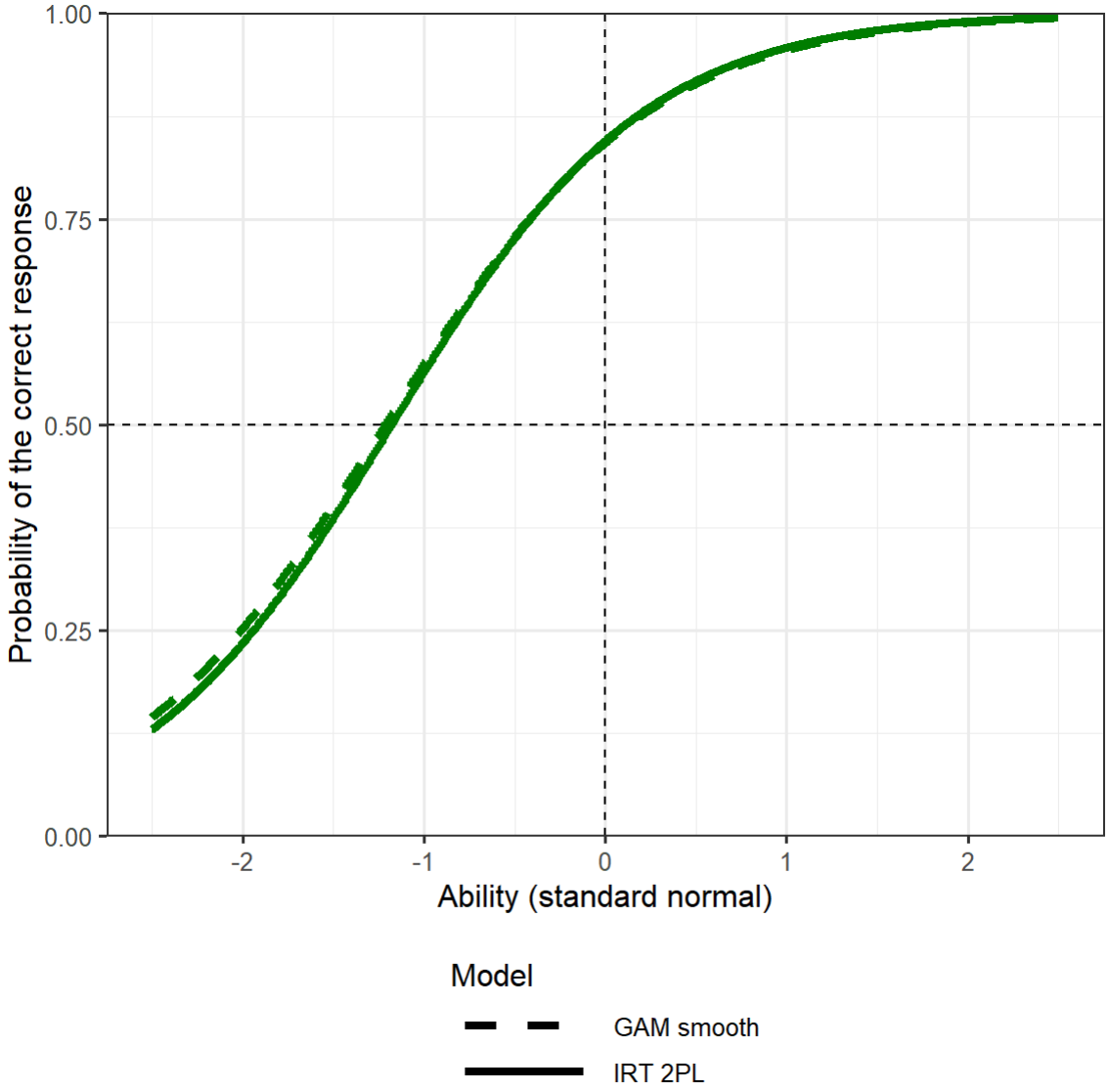
10.2.56.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.2	11.3	16.1	27.1	61.1	2.8	0.6
Middle quartiles	3.9	11.6	14.9	23.5	107.0	2.8	0.5
Upper quartile	6.3	10.6	12.8	17.6	44.8	2.6	0.4
Overall	3.2	11.3	14.9	22.1	107.0	2.8	0.5

10.2.57 S100008 (LLM)

10.2.57.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	78.19%	299	LLM simple setup	MC	0.44	0.25	1.44	−1.18	0.000

10.2.57.2 Gender DIF:

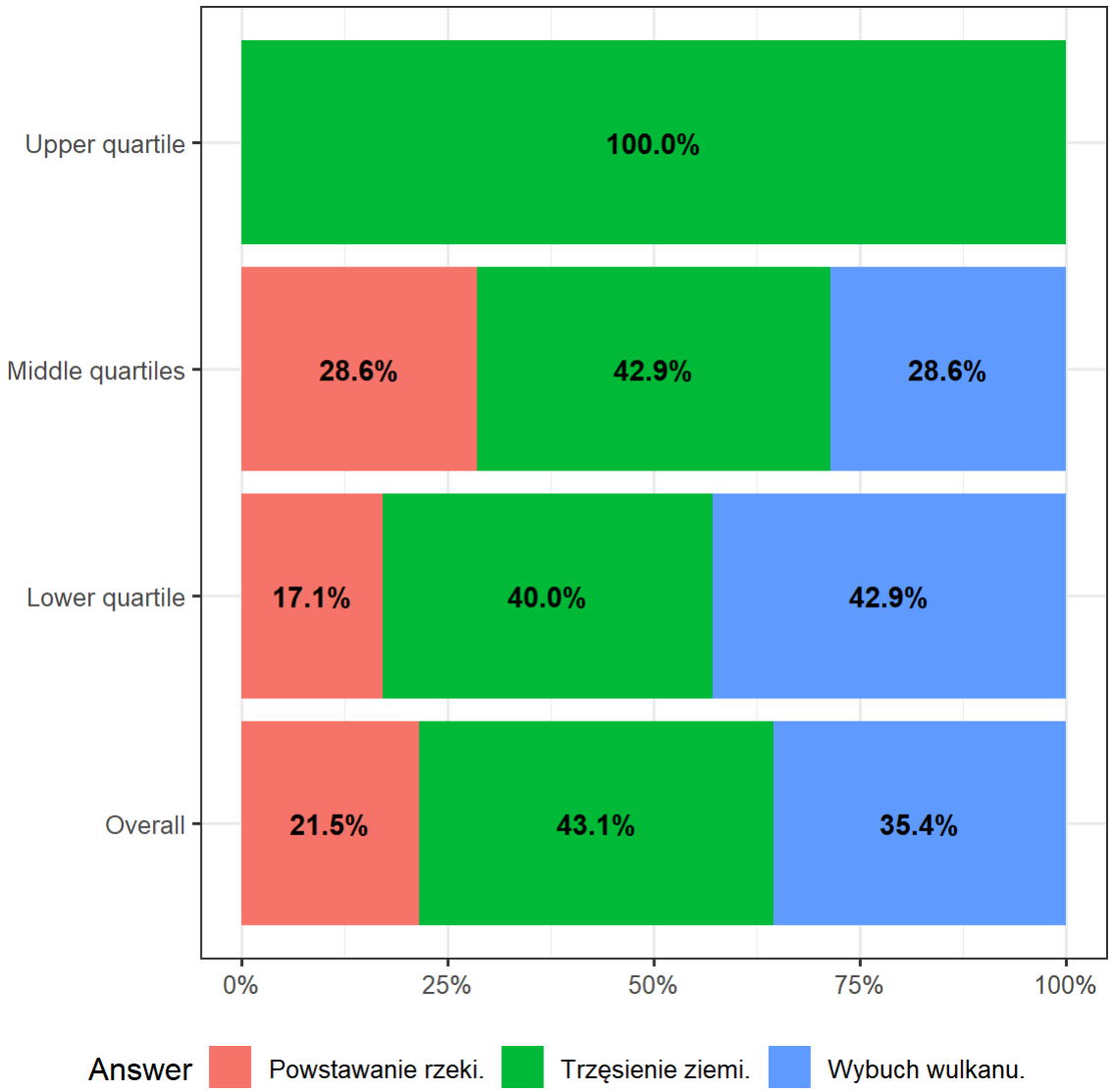
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.00	0.369	1.000	0.007	N/A

10.2.57.3 Frequency of answers:

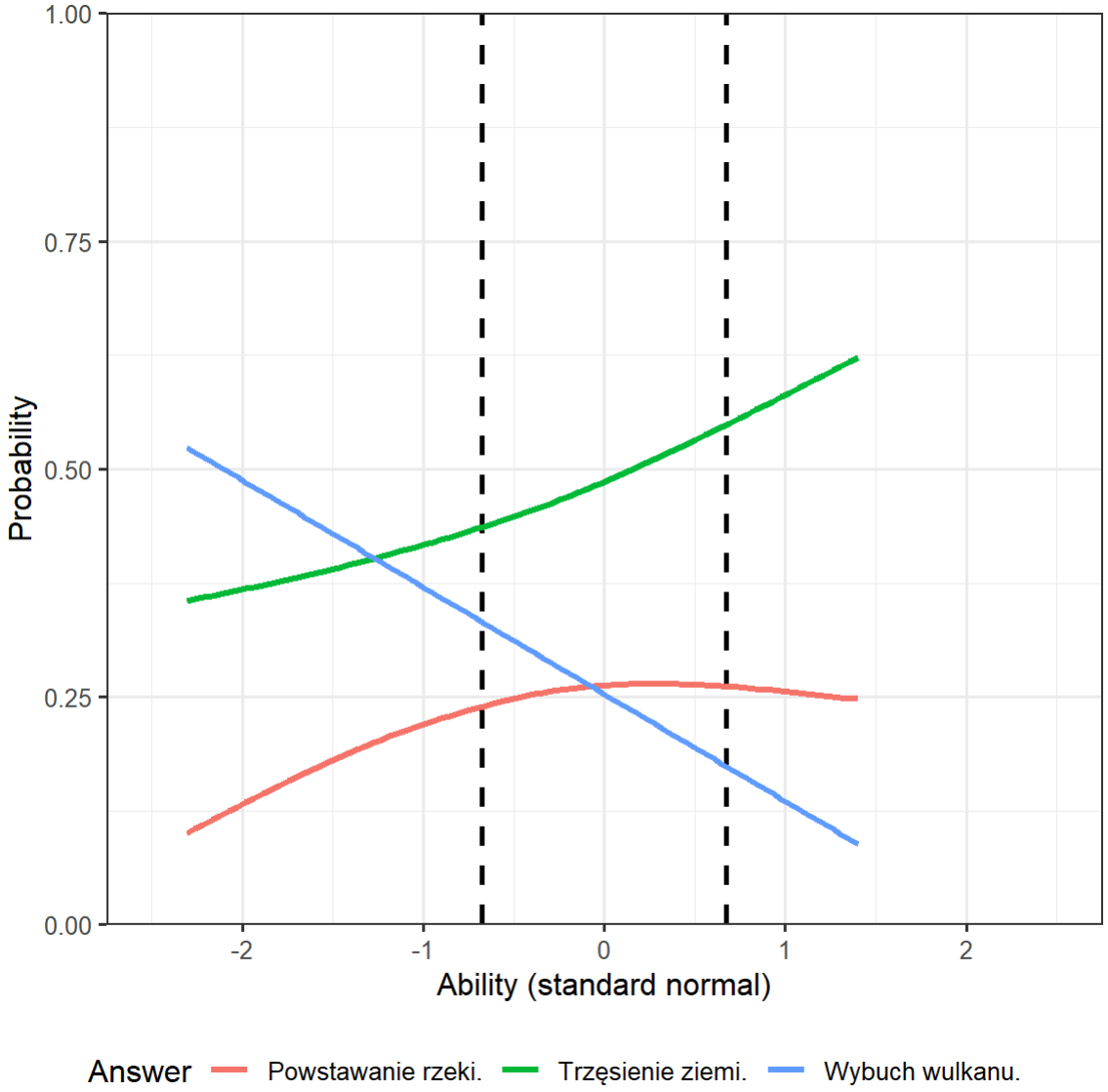
Answer	Frequency	Percent
Powstawanie fal na morzu. (correct)	233	78.2%
Powstawanie rzeki.	14	4.7%
Trzęsienie ziemi.	28	9.4%
Wybuch wulkanu.	23	7.7%
Overall	298	100.0%

10.2.57.4 Distractors functioning:

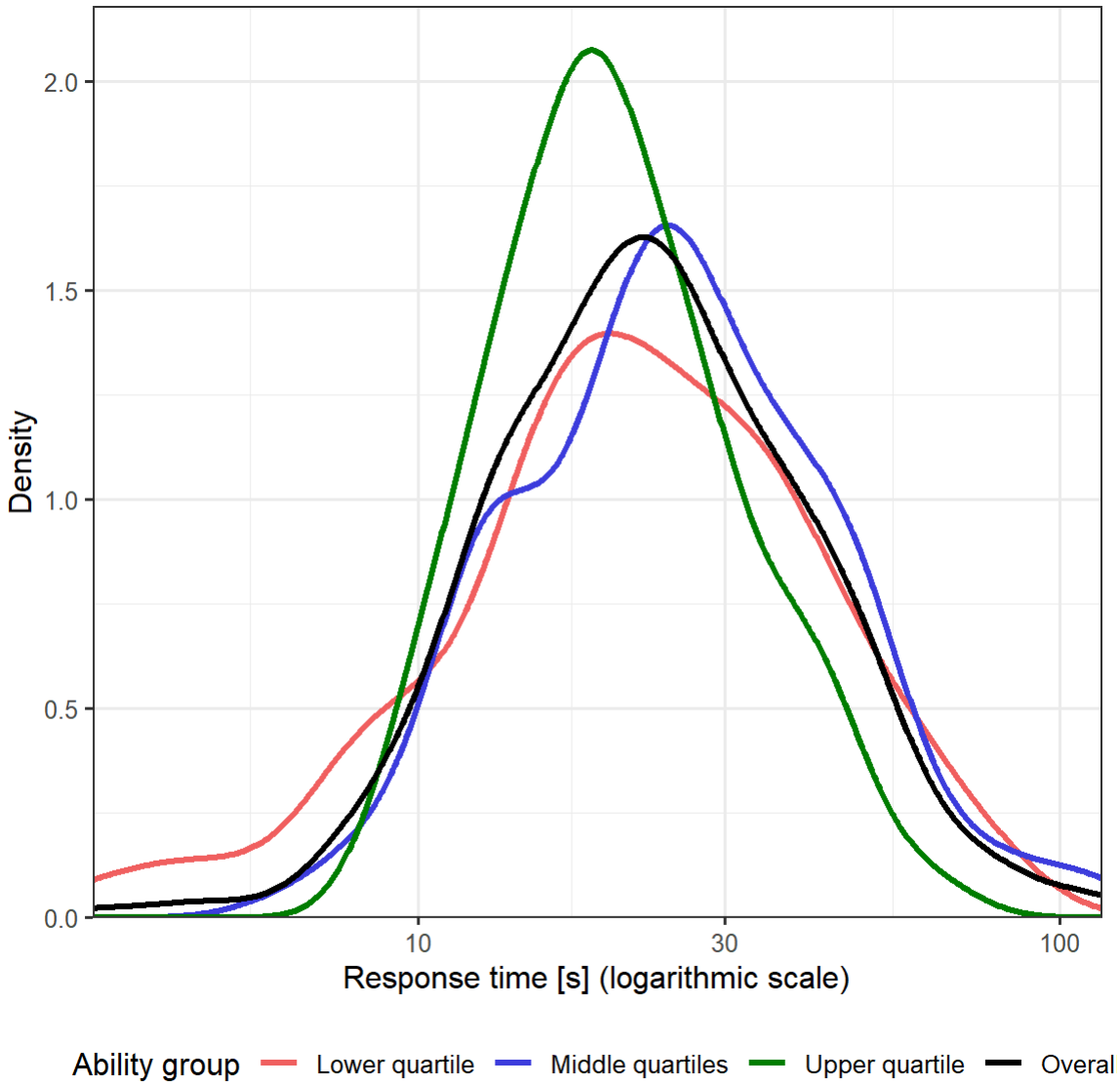
	Frequency				Percent			
Answer	Lower quartile	Middle quartiles	Upper quartile	Overall	Lower quartile	Middle quartiles	Upper quartile	Overall
Powstawanie rzeki.	6	8	0	14	17.1%	28.6%	0.0%	21.5%
Trzęsienie ziemi.	14	12	2	28	40.0%	42.9%	100.0%	43.1%
Wybuch wulkanu.	15	8	0	23	42.9%	28.6%	0.0%	35.4%
Overall	35	28	2	65	100.0%	100.0%	100.0%	100.0%




```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```



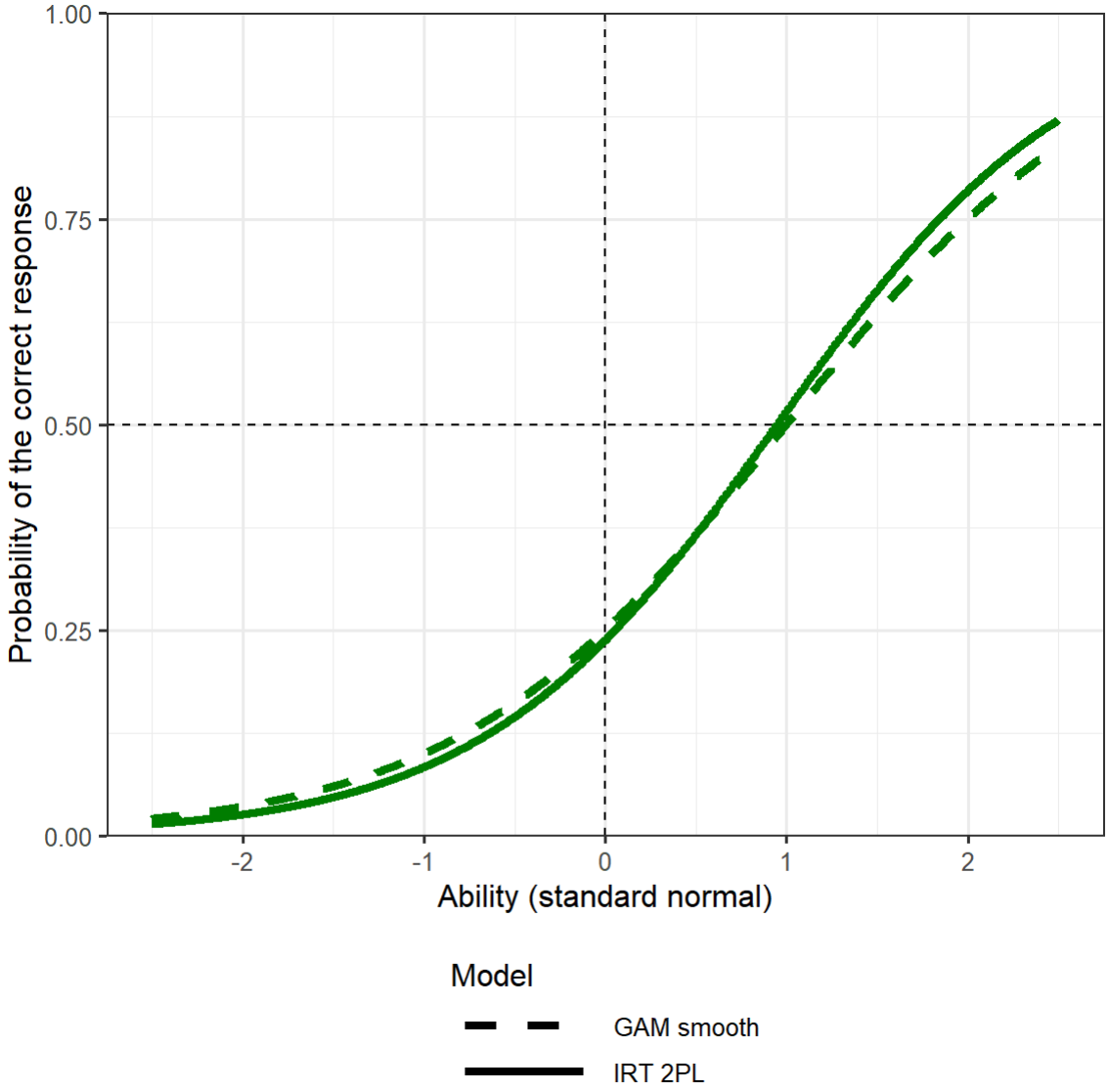
10.2.57.5 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	15.3	21.8	33.3	78.8	3.1	0.7
Middle quartiles	6.2	17.6	24.7	36.4	115.9	3.2	0.6
Upper quartile	9.4	14.8	19.1	27.5	63.2	3.0	0.4
Overall	3.1	15.6	22.9	33.5	115.9	3.1	0.6

10.2.58 S061034 (TIMSS)

10.2.58.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	27.87%	306	TIMSS	CR	0.40	0.46	1.23	0.95	0.000

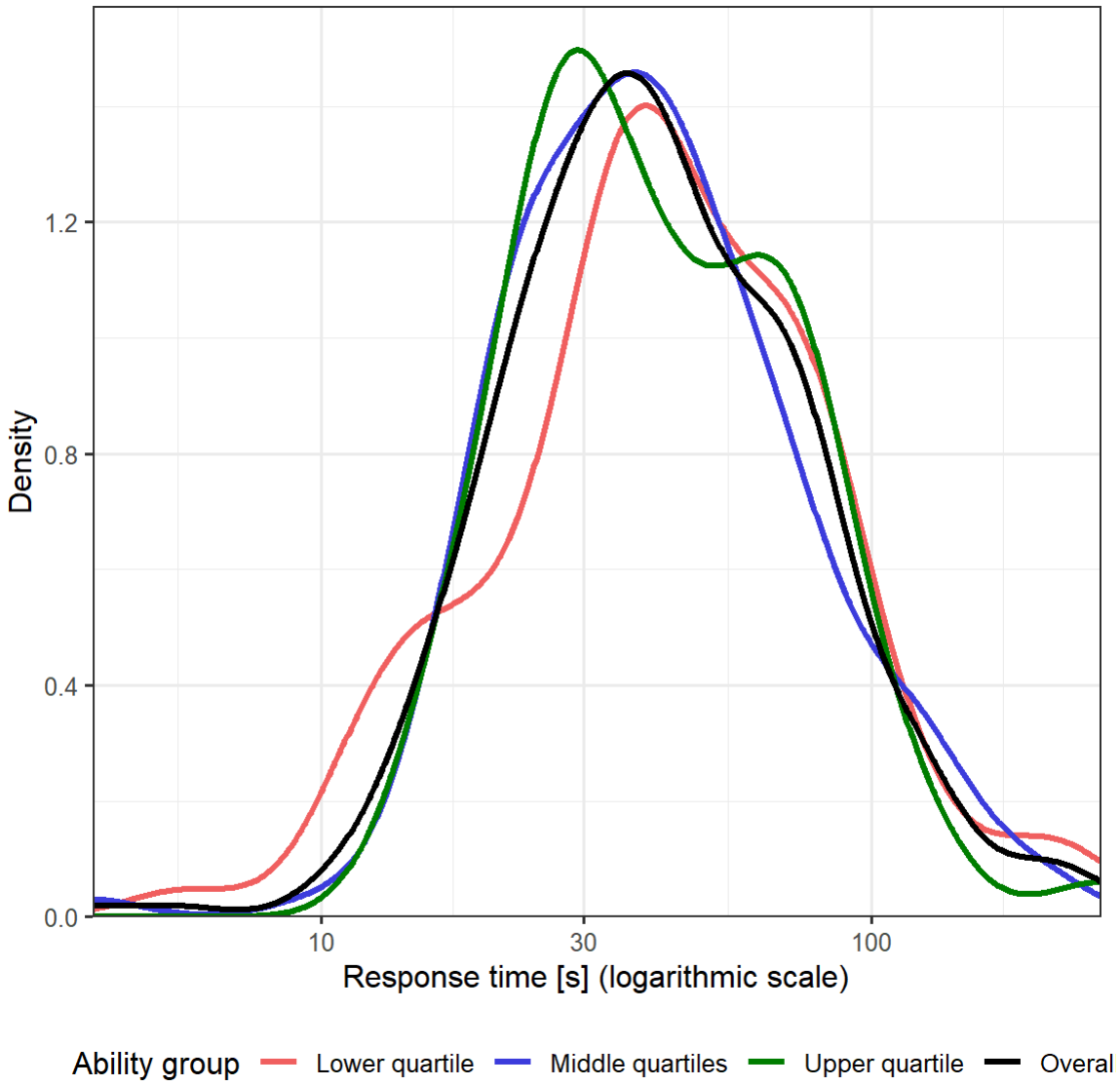
10.2.58.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
4.87	0.088	1.000	0.015	N/A

10.2.58.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

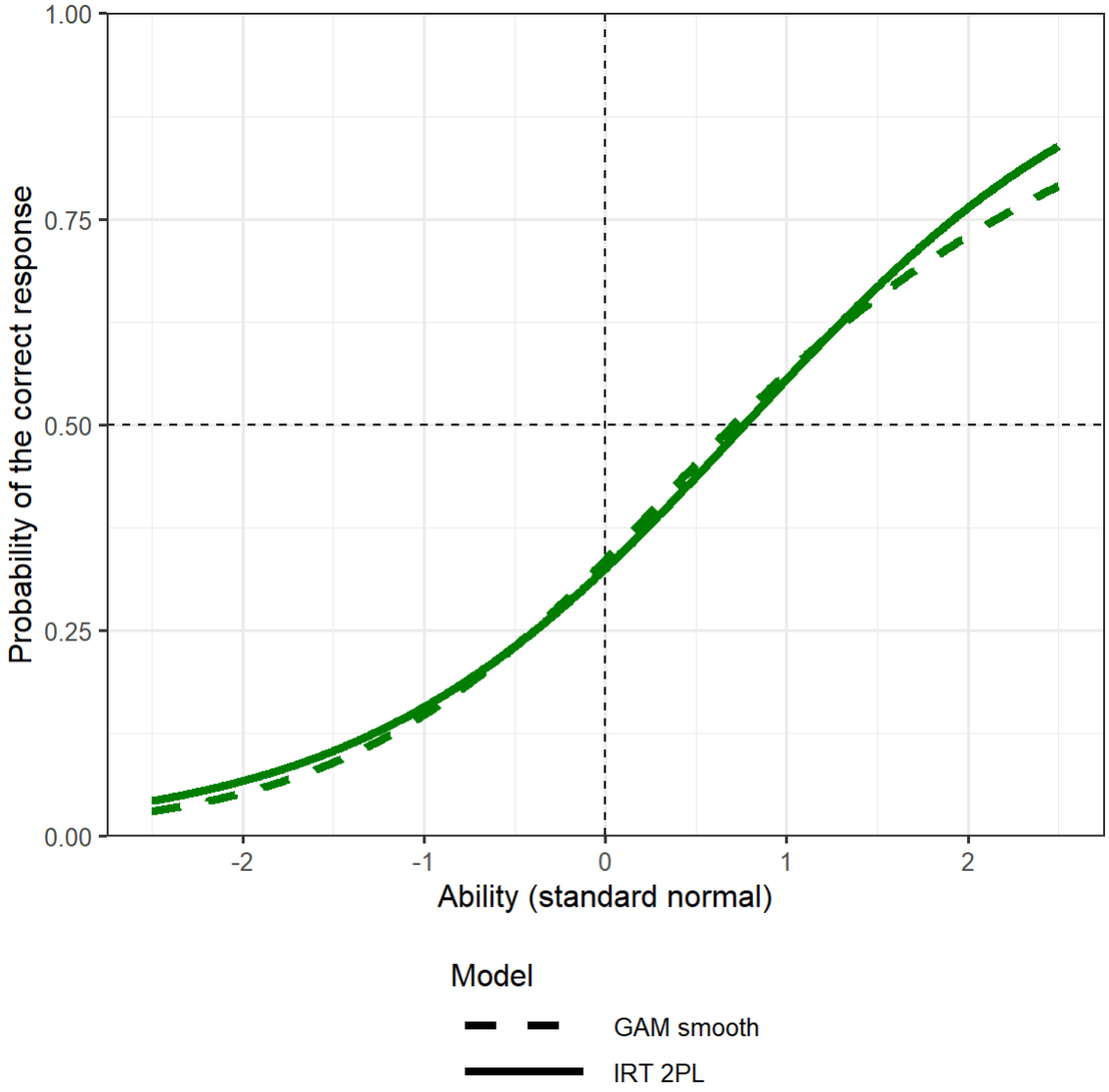
10.2.58.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.6	28.9	40.7	68.4	230.0	3.7	0.7
Middle quartiles	3.8	25.5	38.6	58.5	215.3	3.7	0.6
Upper quartile	14.2	26.5	38.2	65.0	260.0	3.7	0.6
Overall	3.8	26.5	38.9	64.1	260.0	3.7	0.6

10.2.59 S041068 (TIMSS)

10.2.59.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	34.54%	305	TIMSS	CR	0.37	0.46	0.95	0.77	0.016

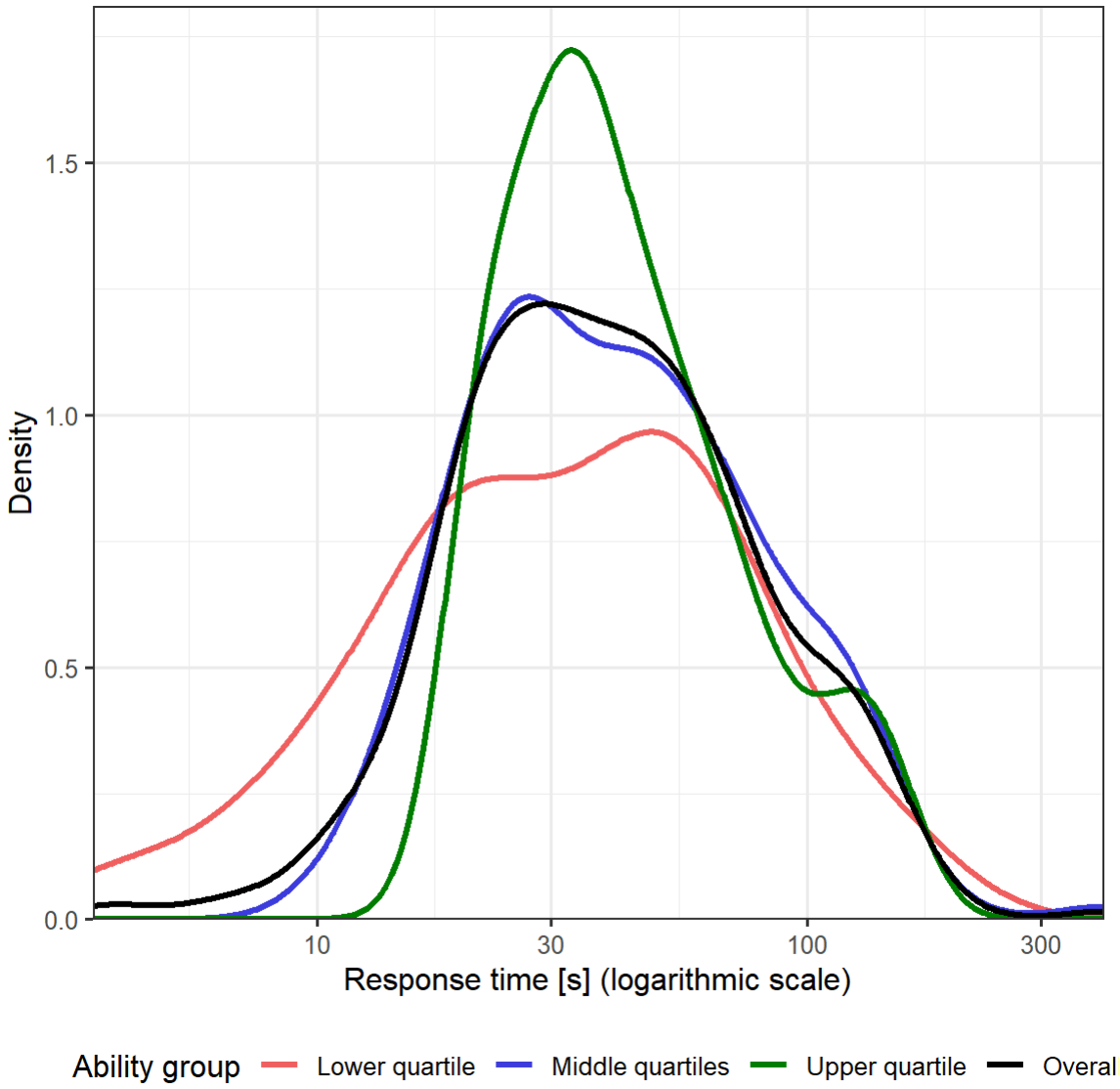
10.2.59.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
3.49	0.174	1.000	0.010	N/A

10.2.59.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

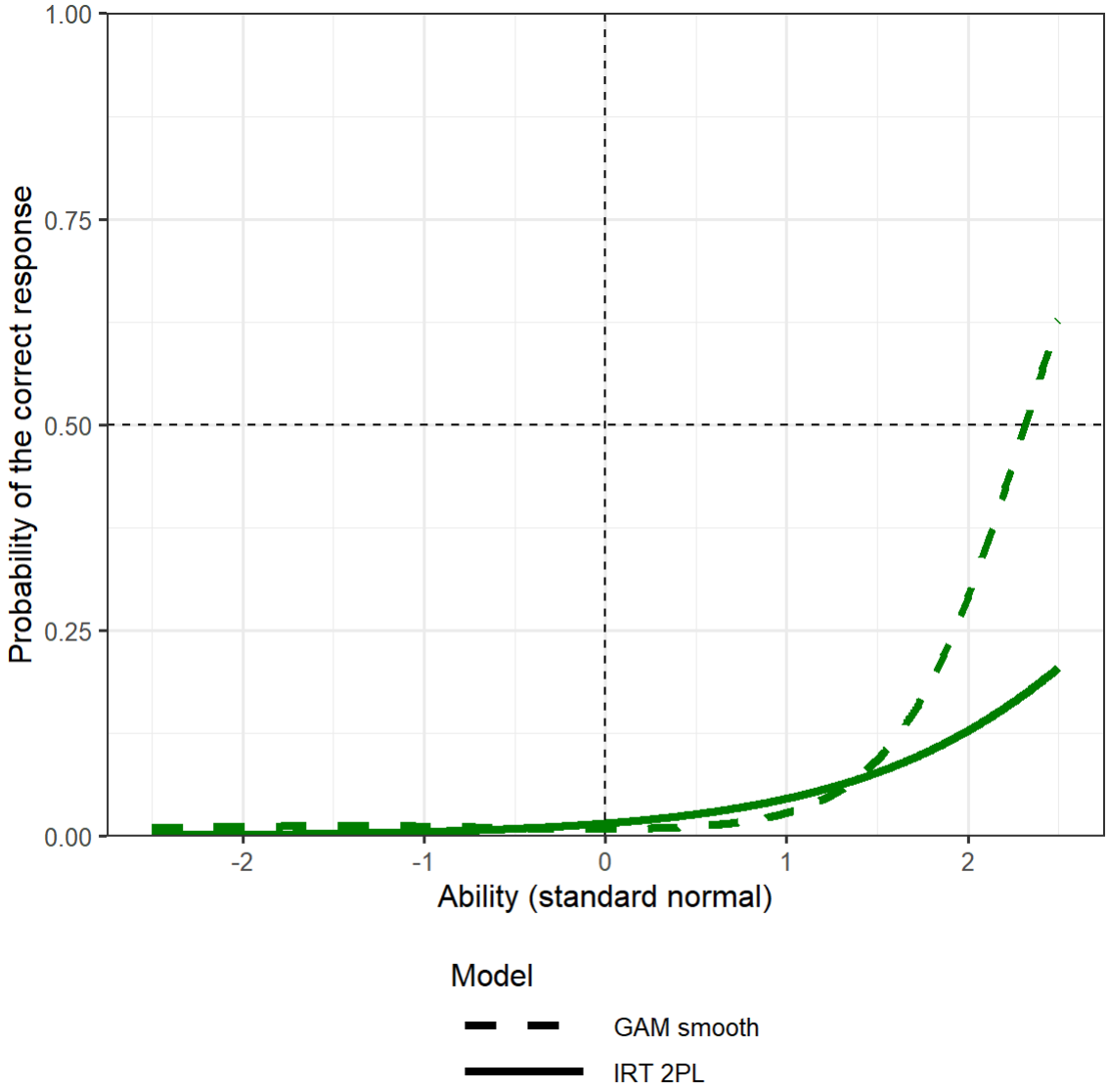
10.2.59.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	18.5	31.9	57.8	170.5	3.4	0.9
Middle quartiles	9.9	24.9	39.2	68.3	401.0	3.7	0.7
Upper quartile	20.3	29.4	38.6	57.5	148.2	3.7	0.5
Overall	3.5	23.9	39.0	62.1	401.0	3.7	0.7

10.2.60 S100004 (LLM)

10.2.60.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	2.69%	299	LLM simple setup	CR	0.19	0.07	1.12	3.71	0.019

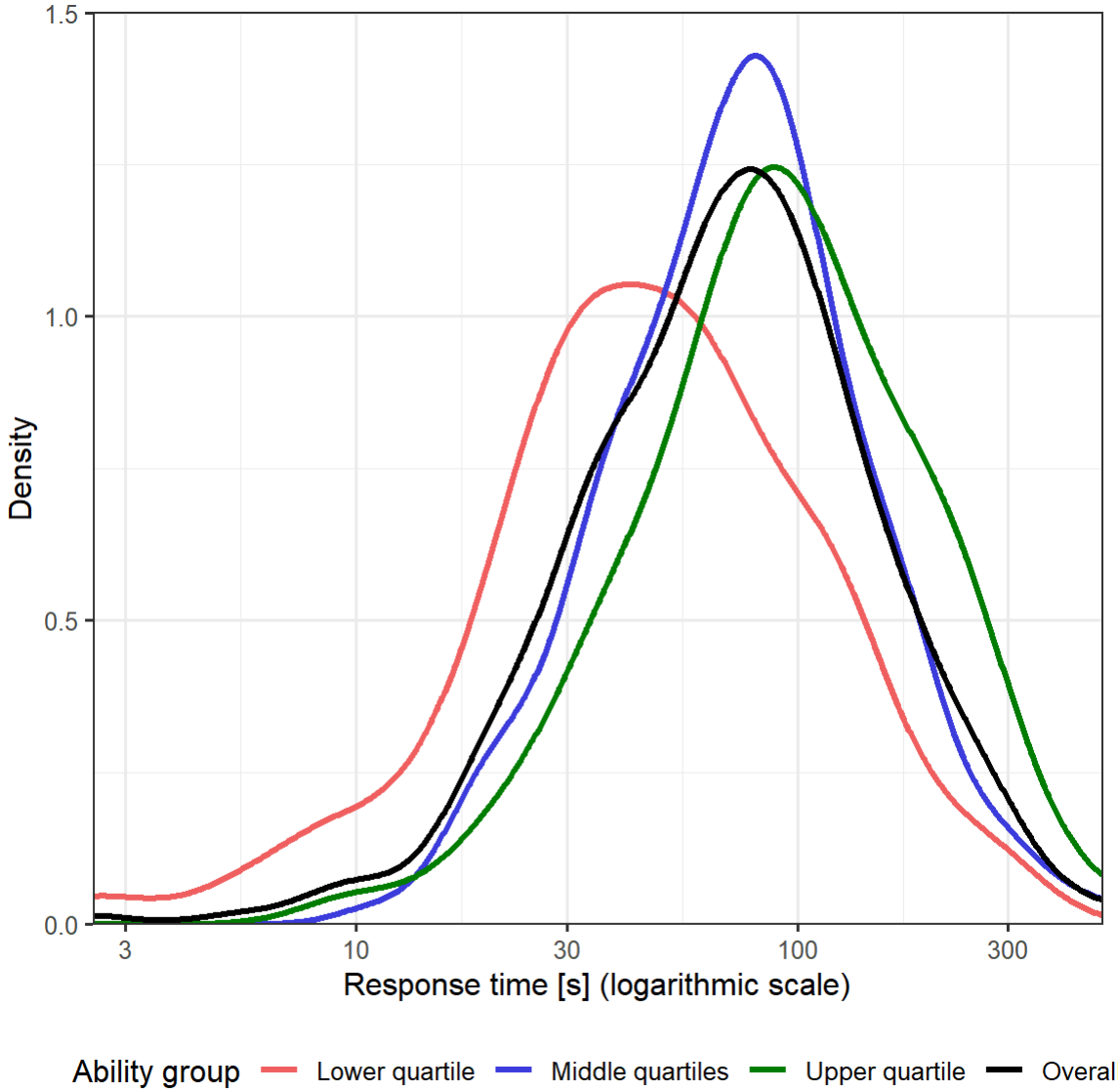
10.2.60.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.14	0.566	1.000	0.016	N/A

10.2.60.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

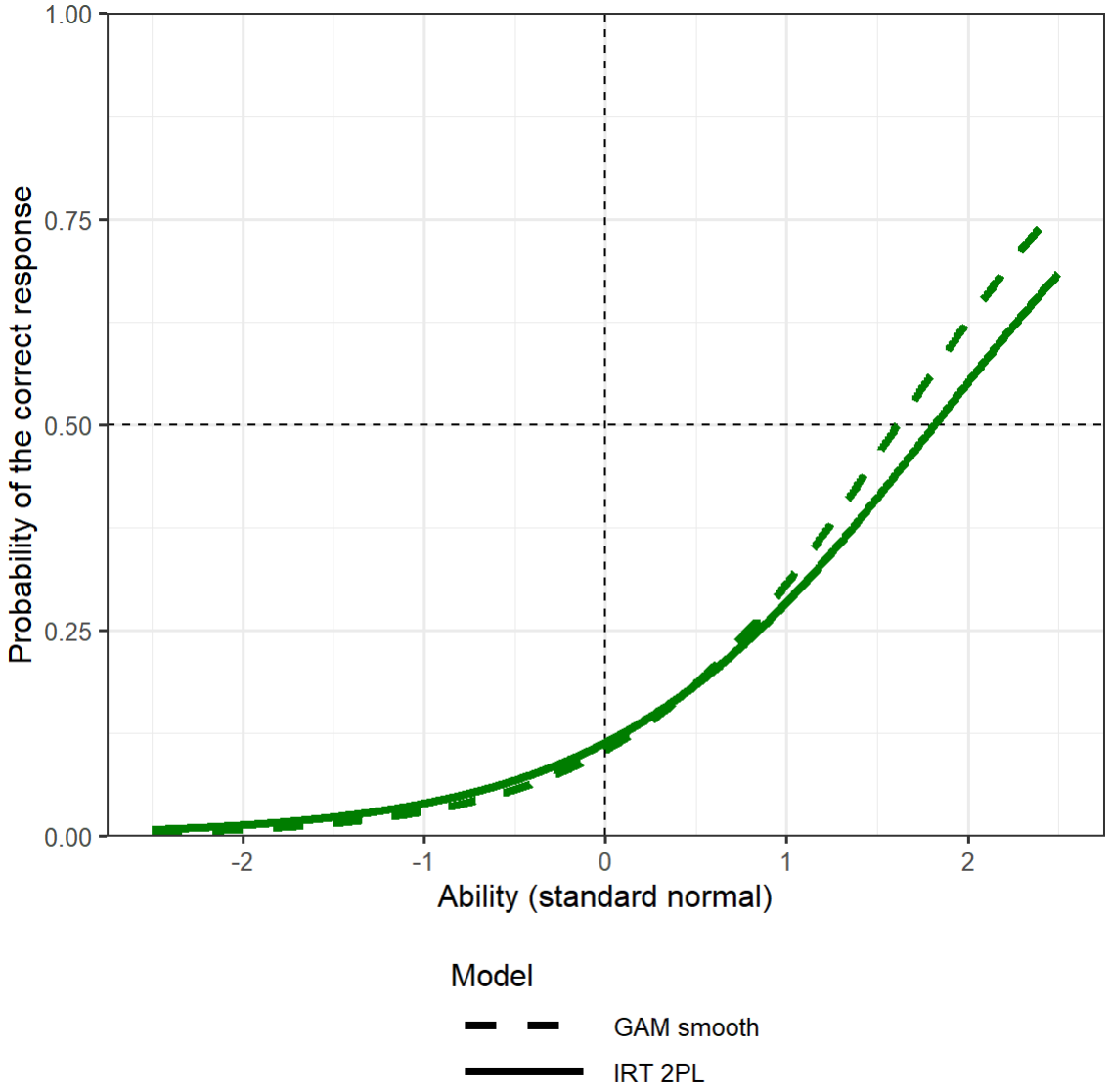
10.2.60.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.5	28.1	48.8	79.4	293.9	3.8	0.9
Middle quartiles	11.2	47.1	75.8	110.7	489.6	4.3	0.7
Upper quartile	10.1	58.1	89.1	157.9	460.0	4.5	0.7
Overall	2.5	40.3	71.9	111.8	489.6	4.2	0.8

10.2.61 S041033 (LLM)

10.2.61.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	16.11%	299	TIMSS	CR	0.38	0.31	1.13	1.82	0.015

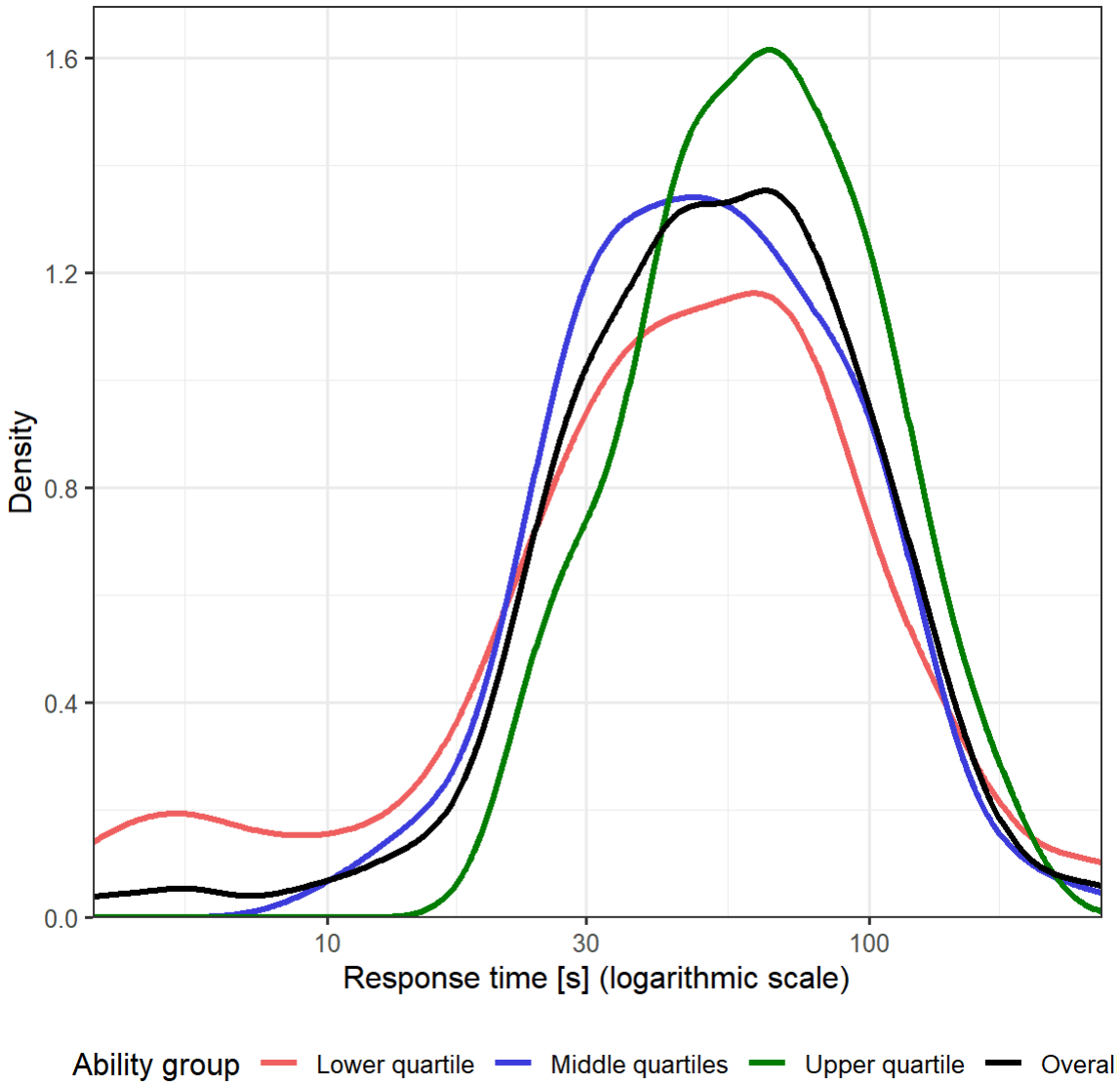
10.2.61.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
11.88	0.003	0.237	0.049	N/A

10.2.61.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

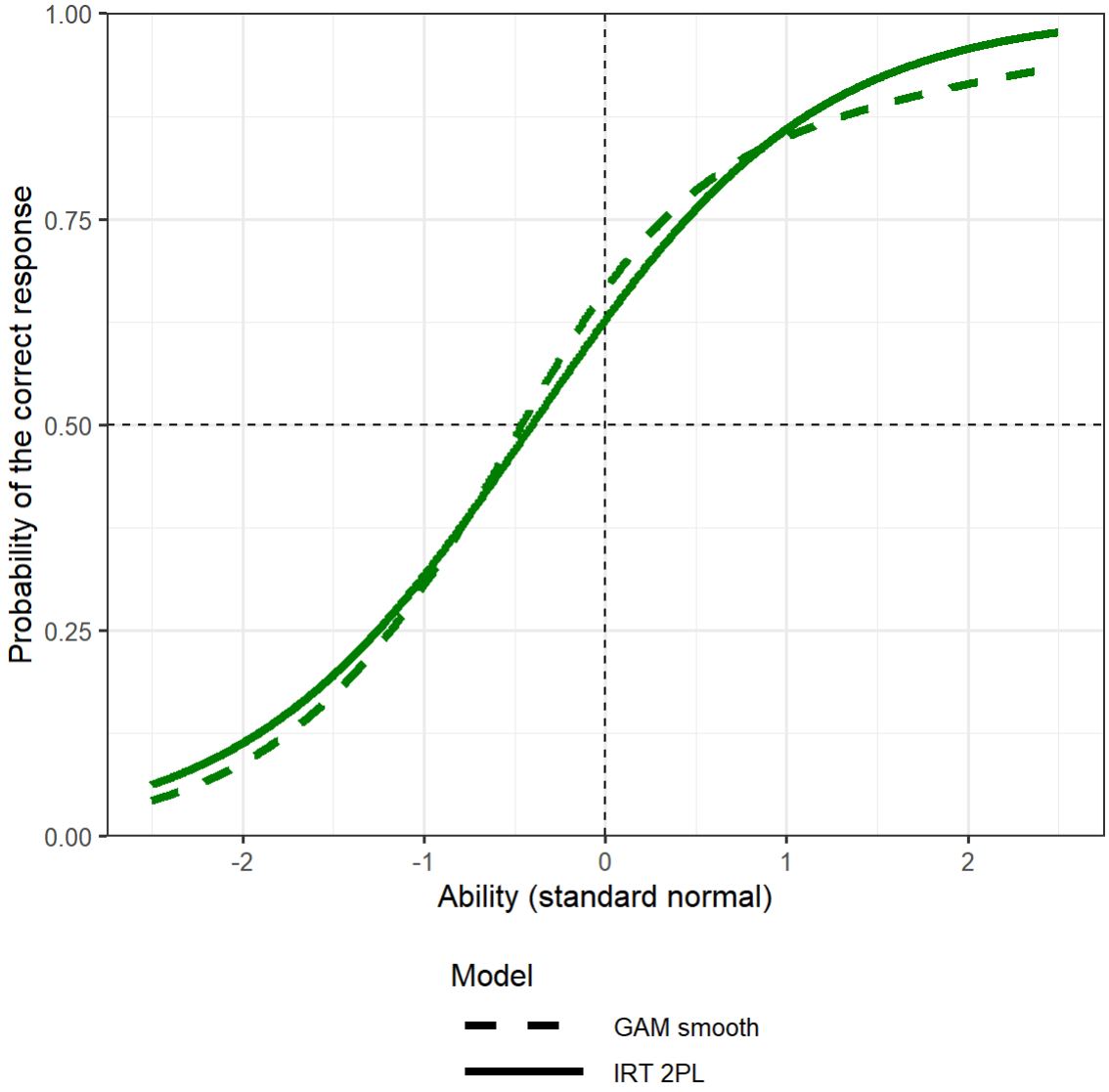
10.2.61.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.7	28.7	43.6	73.6	264.0	3.7	0.9
Middle quartiles	9.8	31.9	48.0	77.4	267.4	3.9	0.6
Upper quartile	22.9	44.0	63.8	90.9	184.8	4.1	0.5
Overall	3.7	33.8	53.0	79.5	267.4	3.9	0.7

10.2.62 S061010 (TIMSS)

10.2.62.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	58.69%	306	TIMSS	CR	0.49	0.47	1.29	−0.40	0.018

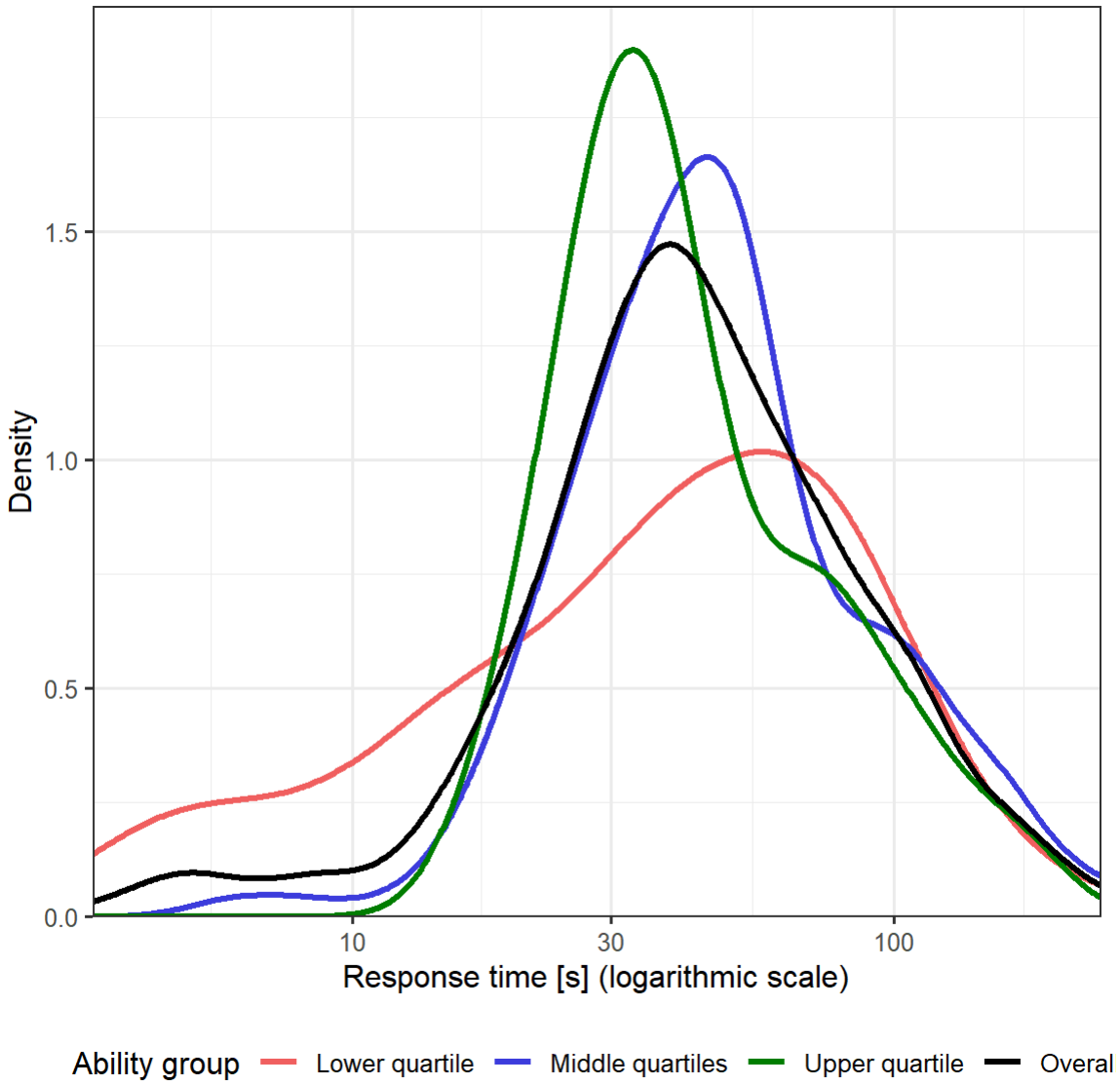
10.2.62.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.80	0.669	1.000	0.002	N/A

10.2.62.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

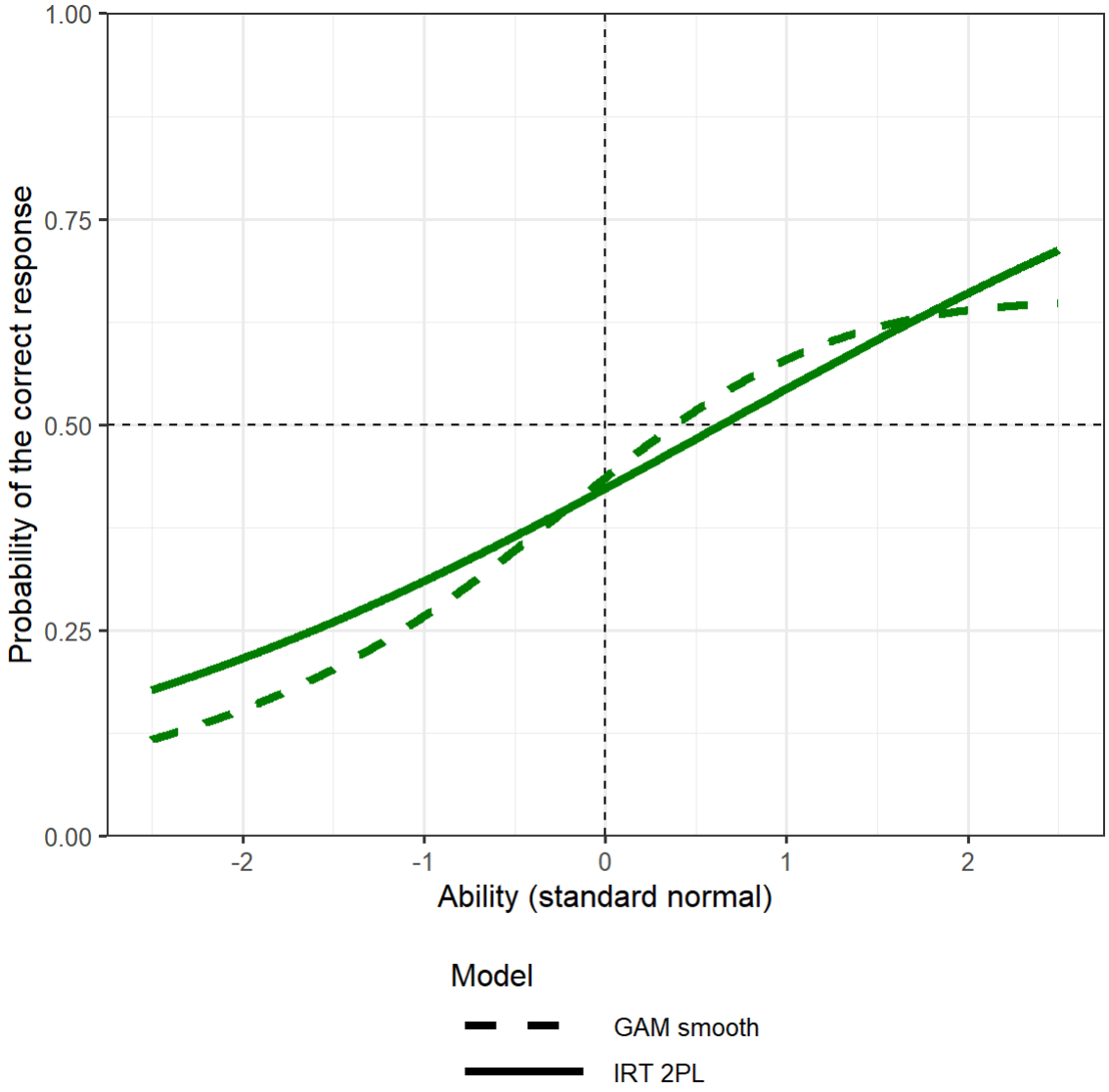
10.2.62.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.3	18.3	39.6	70.3	216.7	3.5	0.9
Middle quartiles	5.9	31.3	46.1	64.9	240.0	3.8	0.6
Upper quartile	16.8	28.6	38.1	61.3	176.2	3.7	0.6
Overall	3.3	28.1	40.8	66.3	240.0	3.7	0.7

10.2.63 S090077 (LLM)

10.2.63.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS generated items: Discrimination	Item difficulty	RMSE
1	42.43%	304	LLM teachers' selection	CR	0.26	0.08	0.49	0.64	0.0(

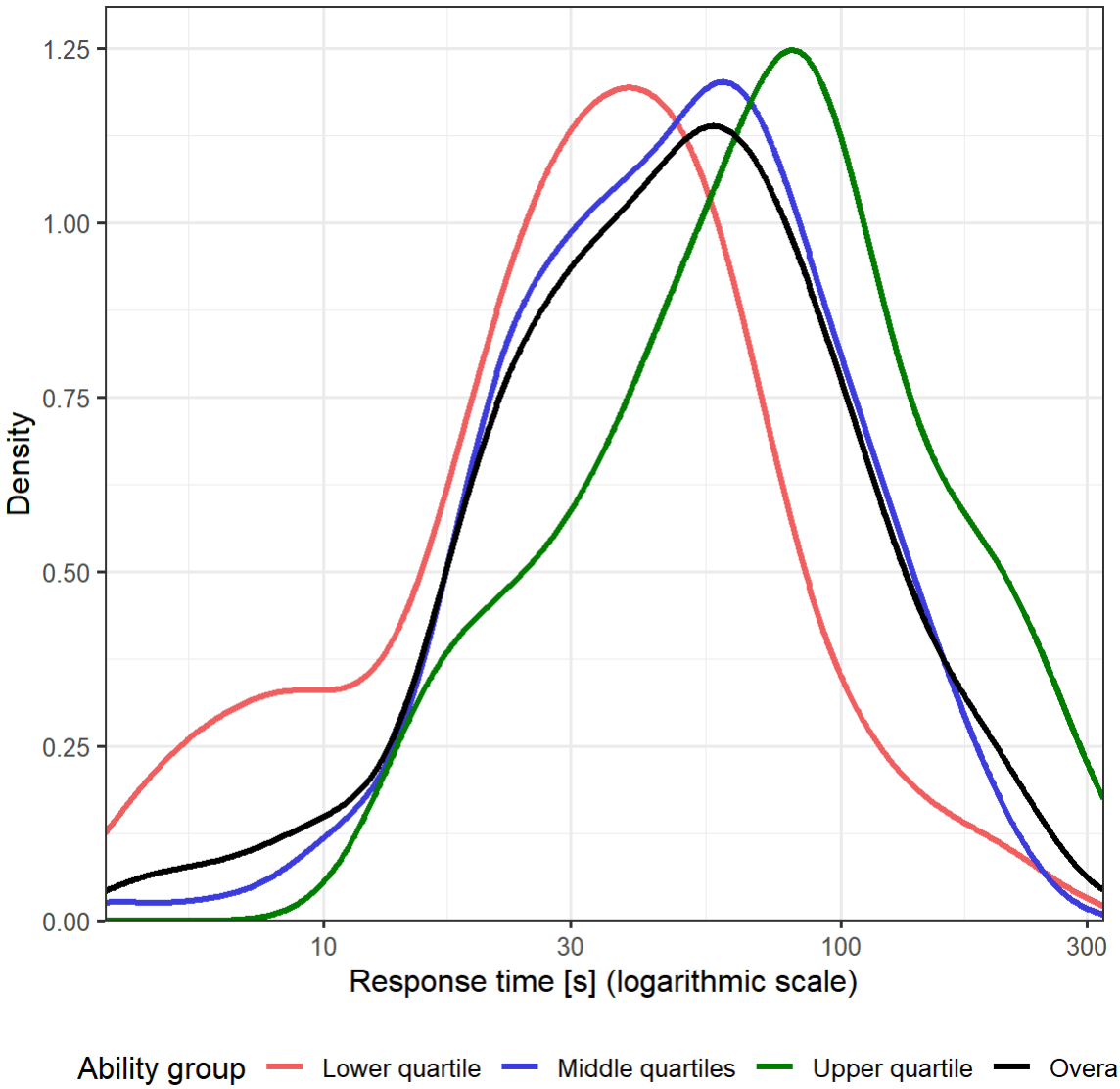
10.2.63.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
8.48	0.014	1.000	0.024	N/A

10.2.63.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

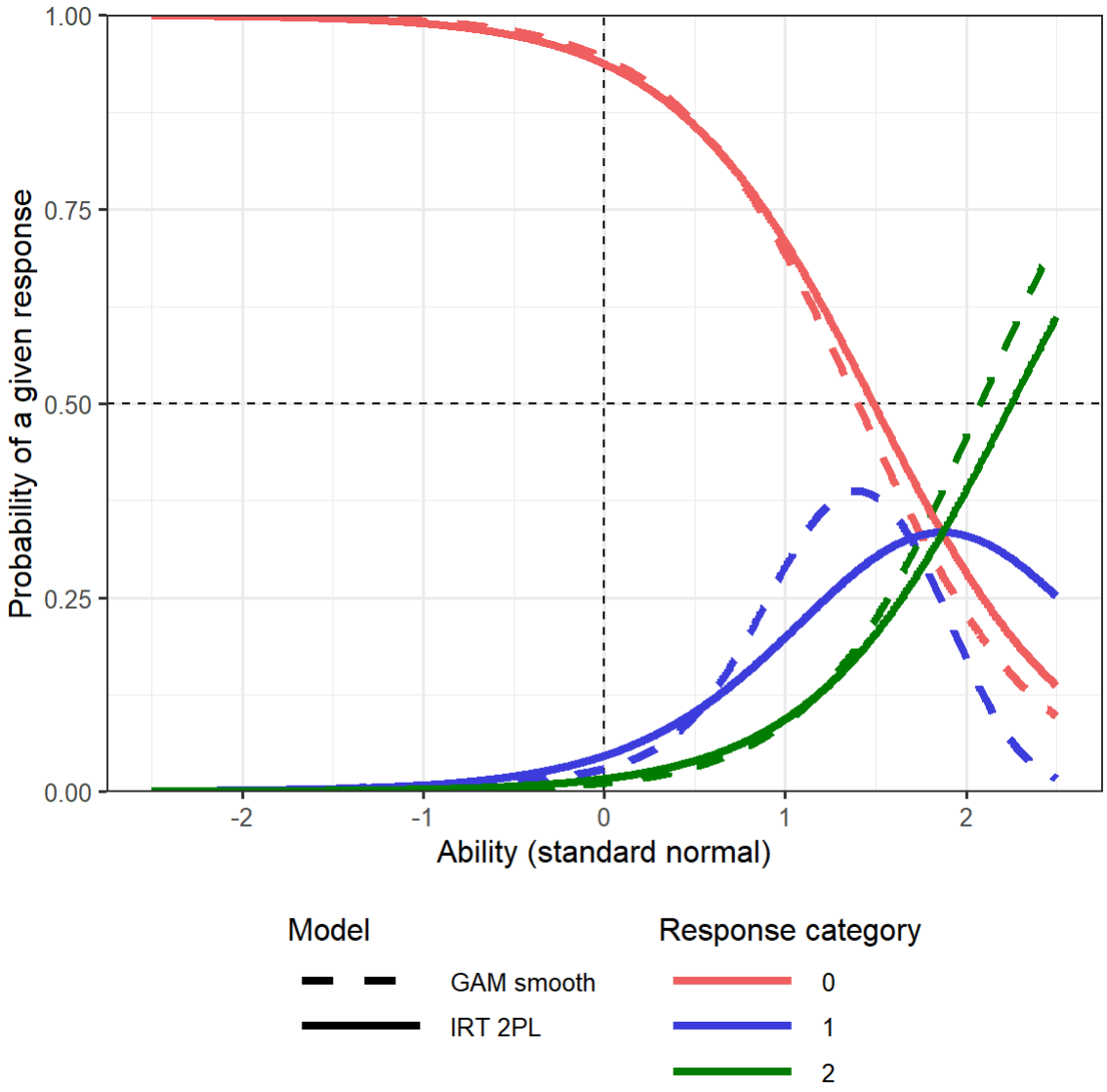
10.2.63.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.6	19.8	33.8	54.3	210.0	3.4	0.8
Middle quartiles	3.8	29.4	50.1	77.1	210.6	3.9	0.7
Upper quartile	14.4	42.2	74.5	105.9	321.9	4.2	0.8
Overall	3.8	28.0	49.9	79.7	321.9	3.8	0.8

10.2.64 S100039 (LLM)

10.2.64.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMS Discrimination	IRT model including both TIMS Item difficulty	IRT model including both TIMS Catego difficul
2	9.57%	304	LLM simple setup	CR	0.49	0.21	1.82	1.87	1.4

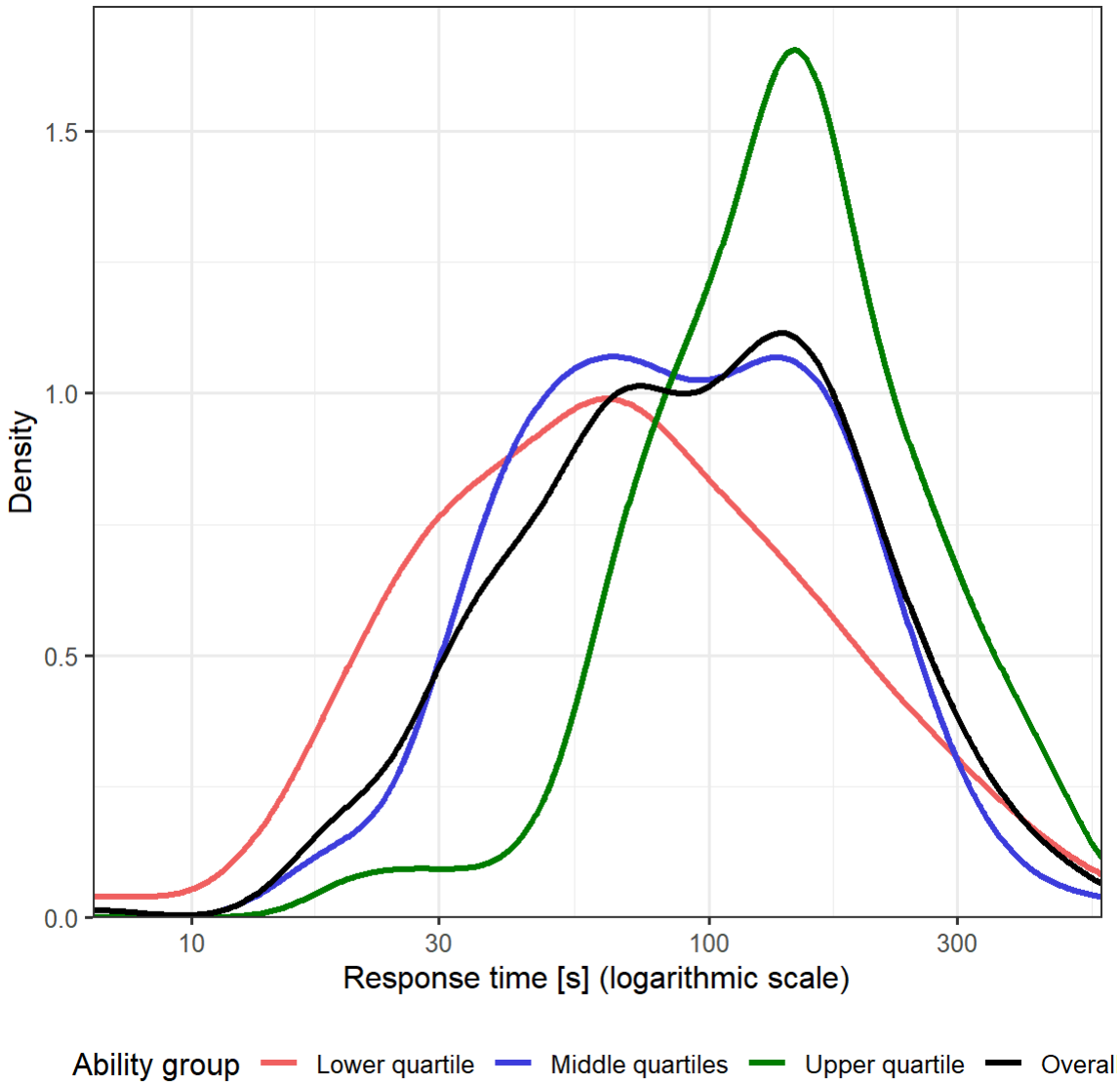
10.2.64.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.30	0.859	1.000	0.001	N/A

10.2.64.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

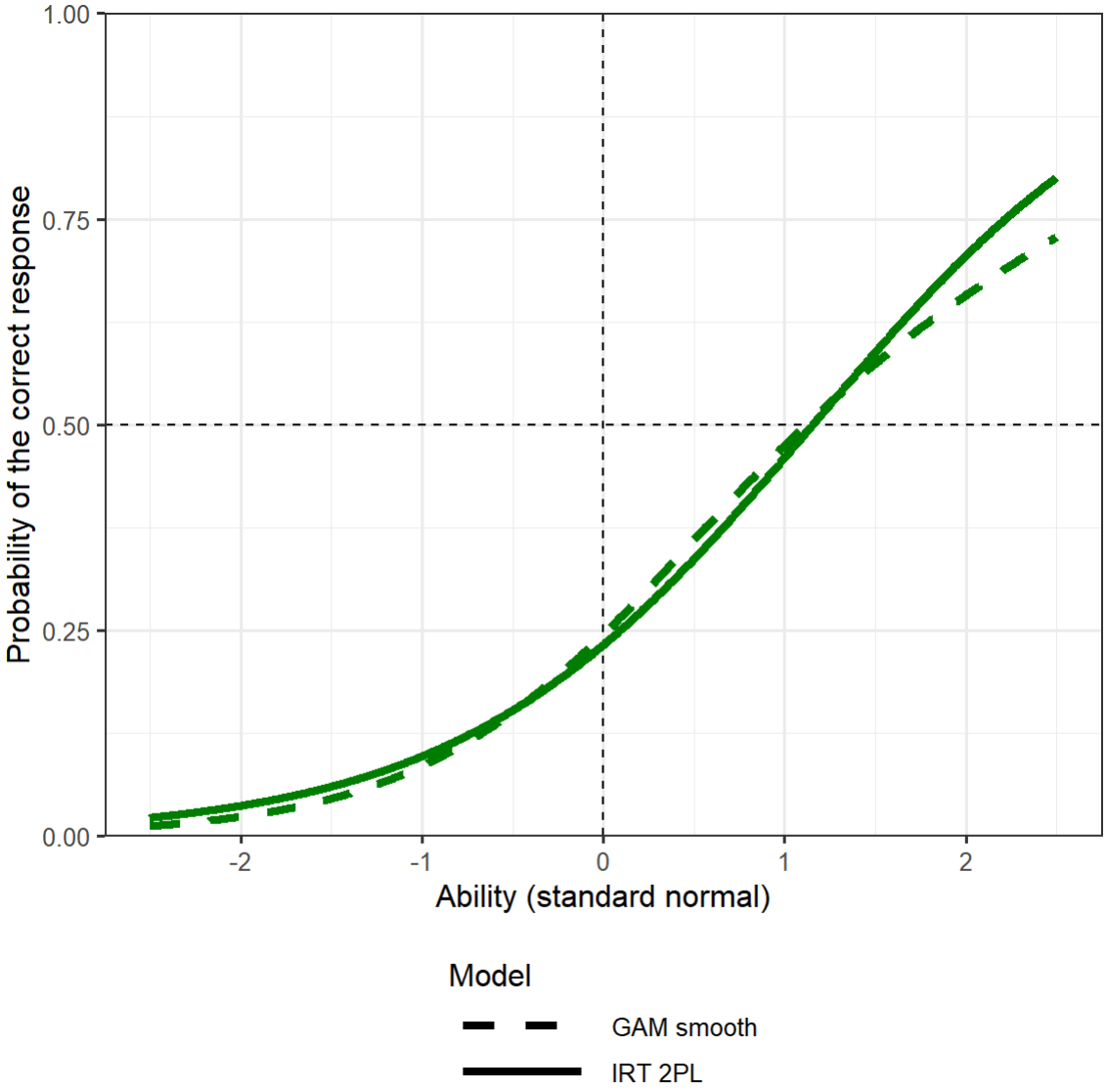
10.2.64.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	6.4	34.3	64.8	119.9	542.7	4.2	0.9
Middle quartiles	16.8	51.1	86.1	148.4	570.1	4.5	0.7
Upper quartile	20.8	95.6	144.8	198.1	483.8	4.9	0.6
Overall	6.4	53.4	95.3	156.8	570.1	4.5	0.8

10.2.65 SP71054 (TIMSS)

10.2.65.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	26.64%	305	TIMSS	CR	0.37	0.42	1.03	1.16	0.007

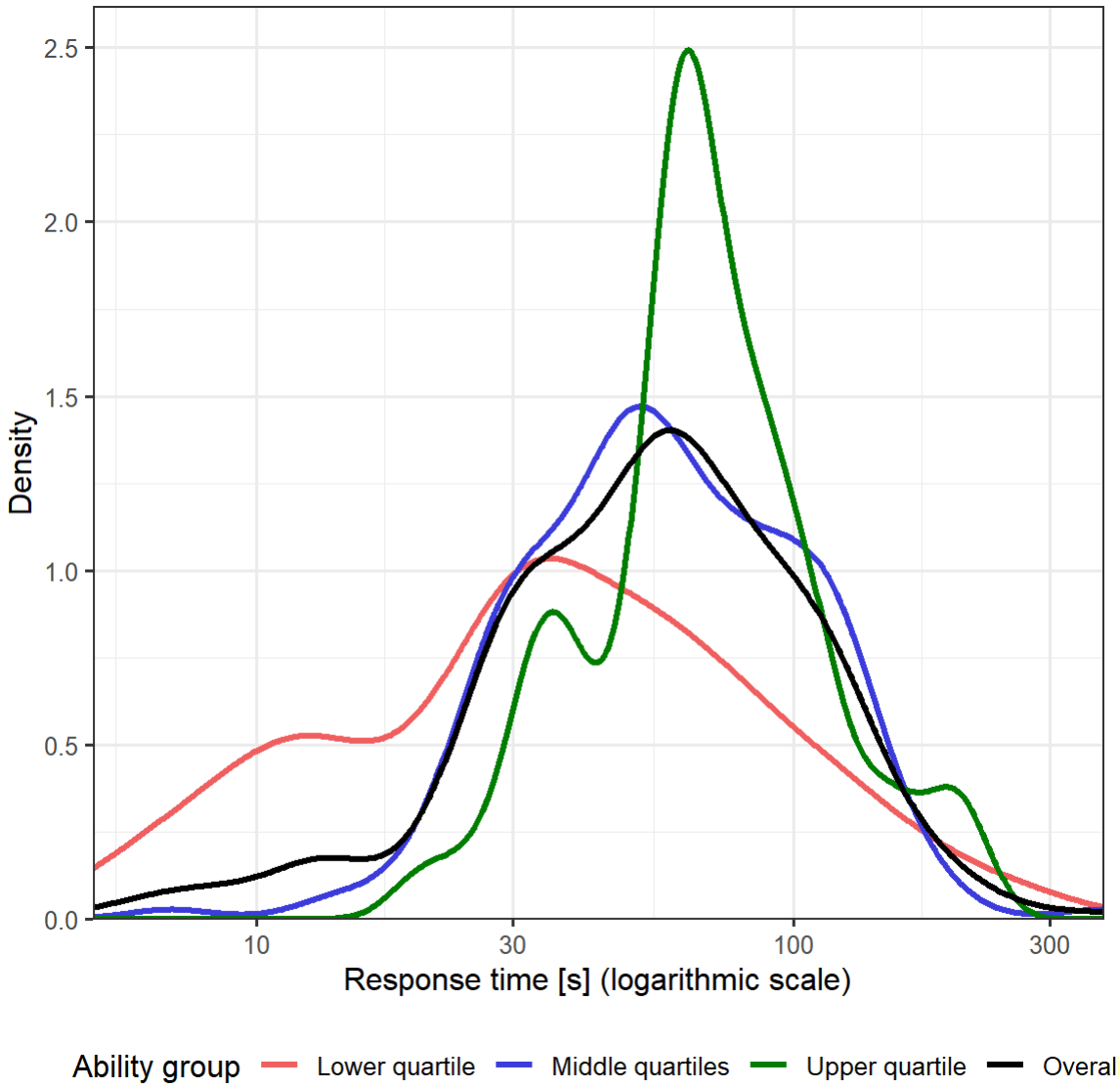
10.2.65.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.93	0.382	1.000	0.006	N/A

10.2.65.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

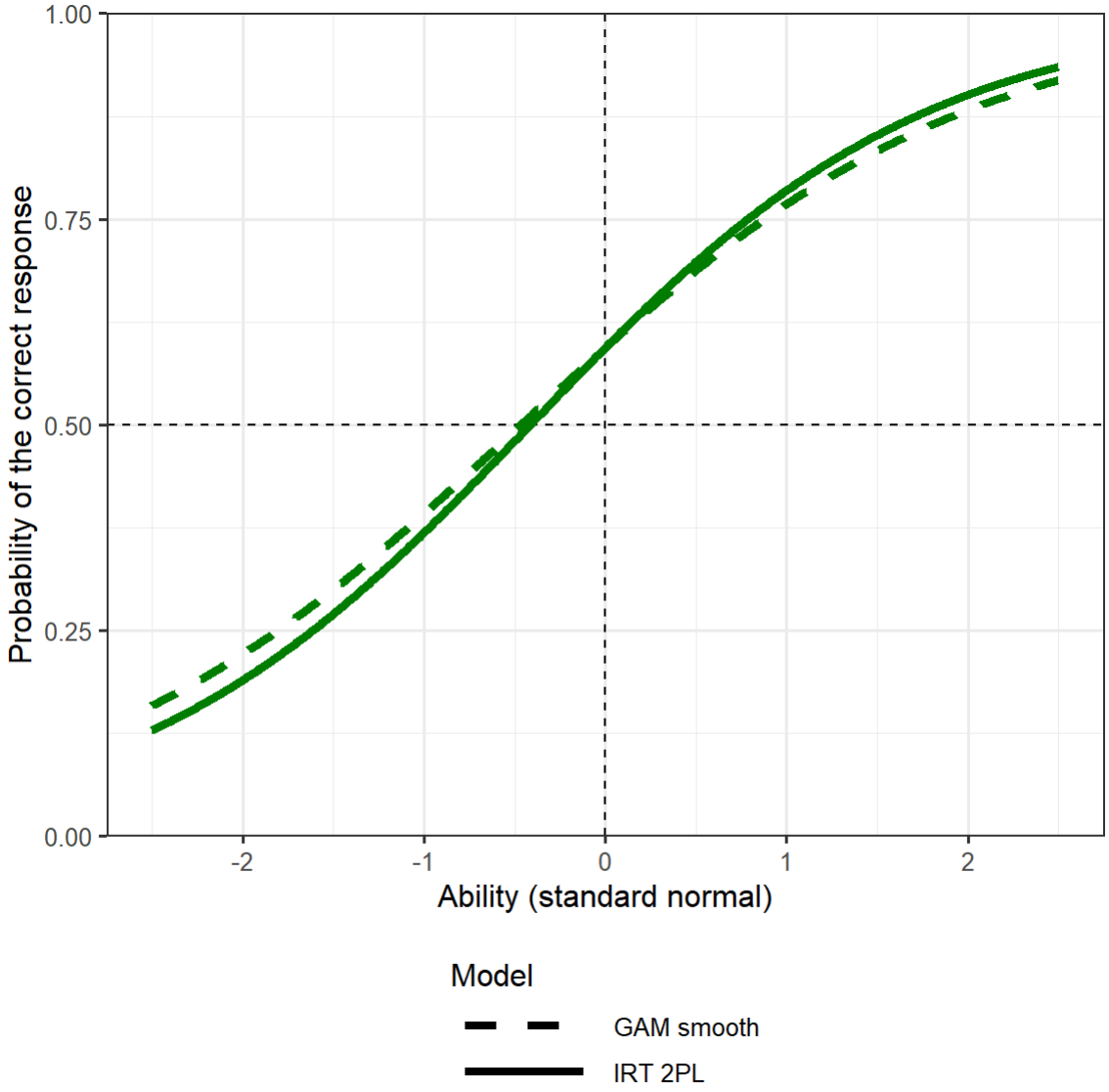
10.2.65.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.0	21.7	35.3	62.0	273.9	3.6	0.9
Middle quartiles	7.0	38.3	55.7	87.4	377.0	4.0	0.6
Upper quartile	20.1	56.3	65.8	88.0	209.1	4.2	0.5
Overall	5.0	34.3	56.2	86.3	377.0	4.0	0.7

10.2.66 S061138 (LLM)

10.2.66.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	57.10%	306	TIMSS	CR	0.35	0.43	0.92	−0.41	0.004

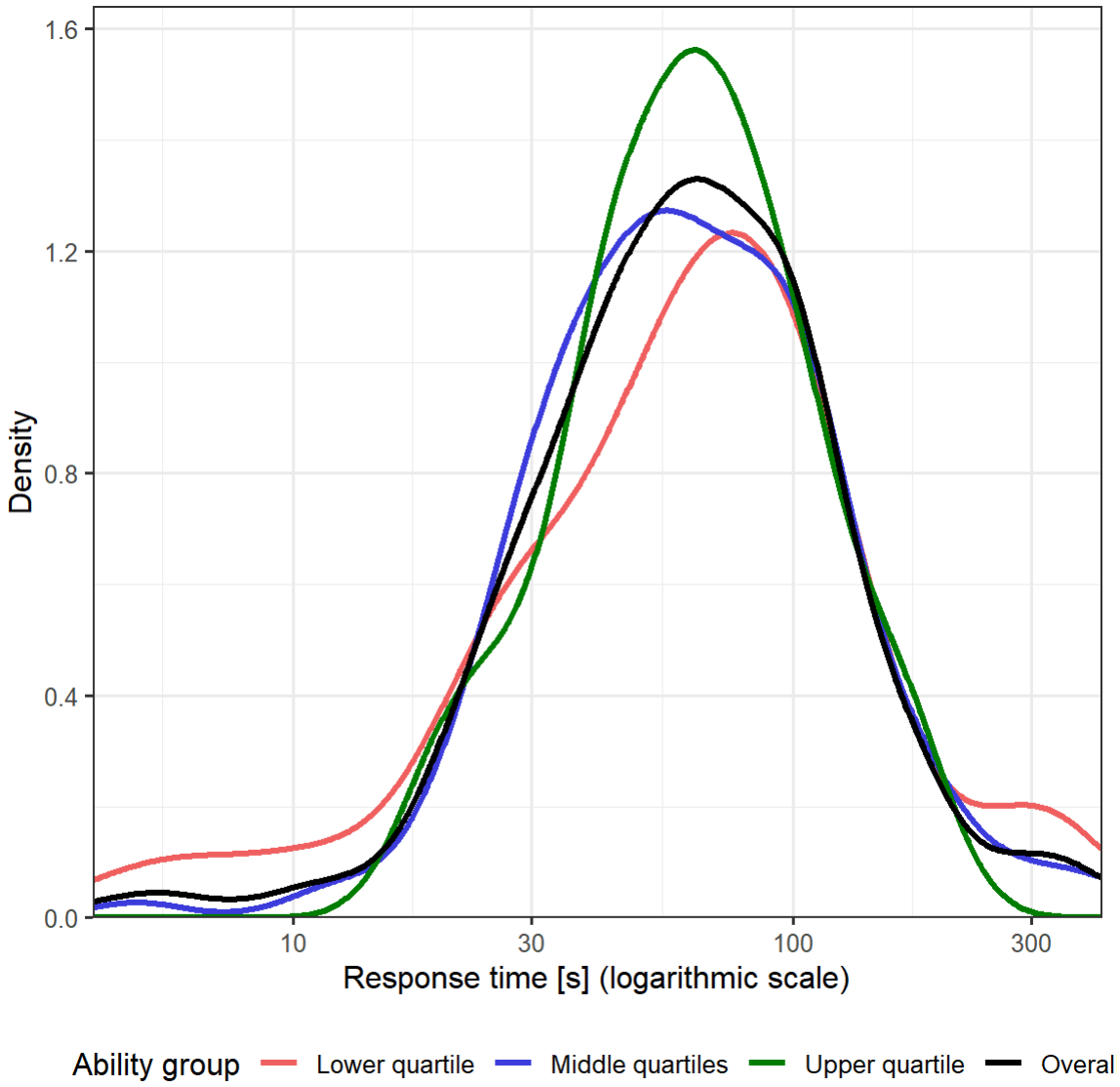
10.2.66.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
10.12	0.006	0.552	0.028	N/A

10.2.66.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

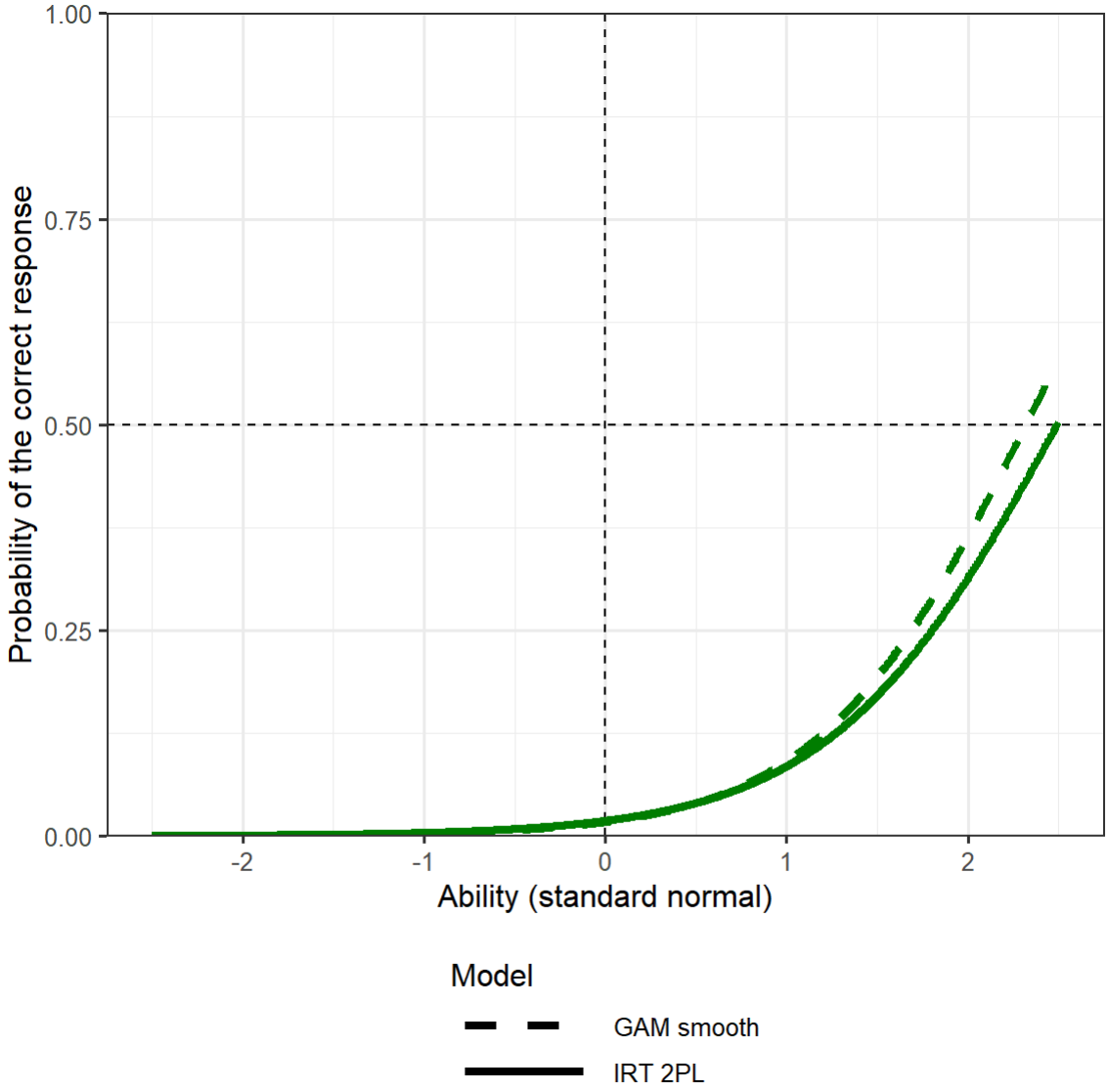
10.2.66.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.0	36.9	61.9	98.2	365.0	4.1	0.9
Middle quartiles	4.9	38.5	60.5	98.0	413.0	4.1	0.7
Upper quartile	18.3	44.0	63.6	92.6	188.2	4.1	0.6
Overall	4.0	38.9	61.9	97.0	413.0	4.1	0.7

10.2.67 S100032 (LLM)

10.2.67.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS a generated items: Discrimination	Item difficulty	RMSEA
1	4.73%	319	LLM simple setup	CR	0.28	0.15	1.59	2.49	0.000

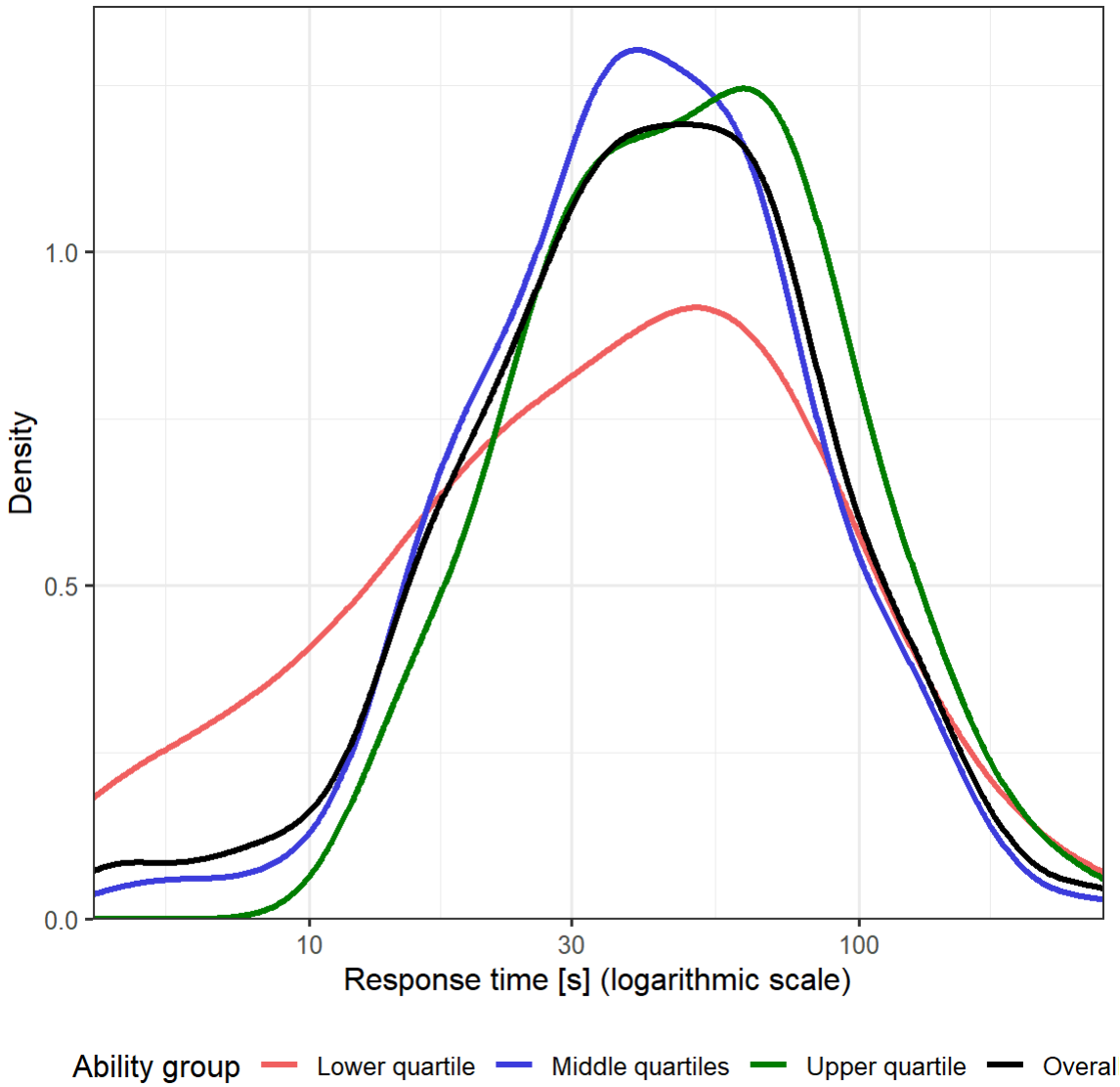
10.2.67.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
12.31	0.002	0.193	0.105	N/A

10.2.67.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

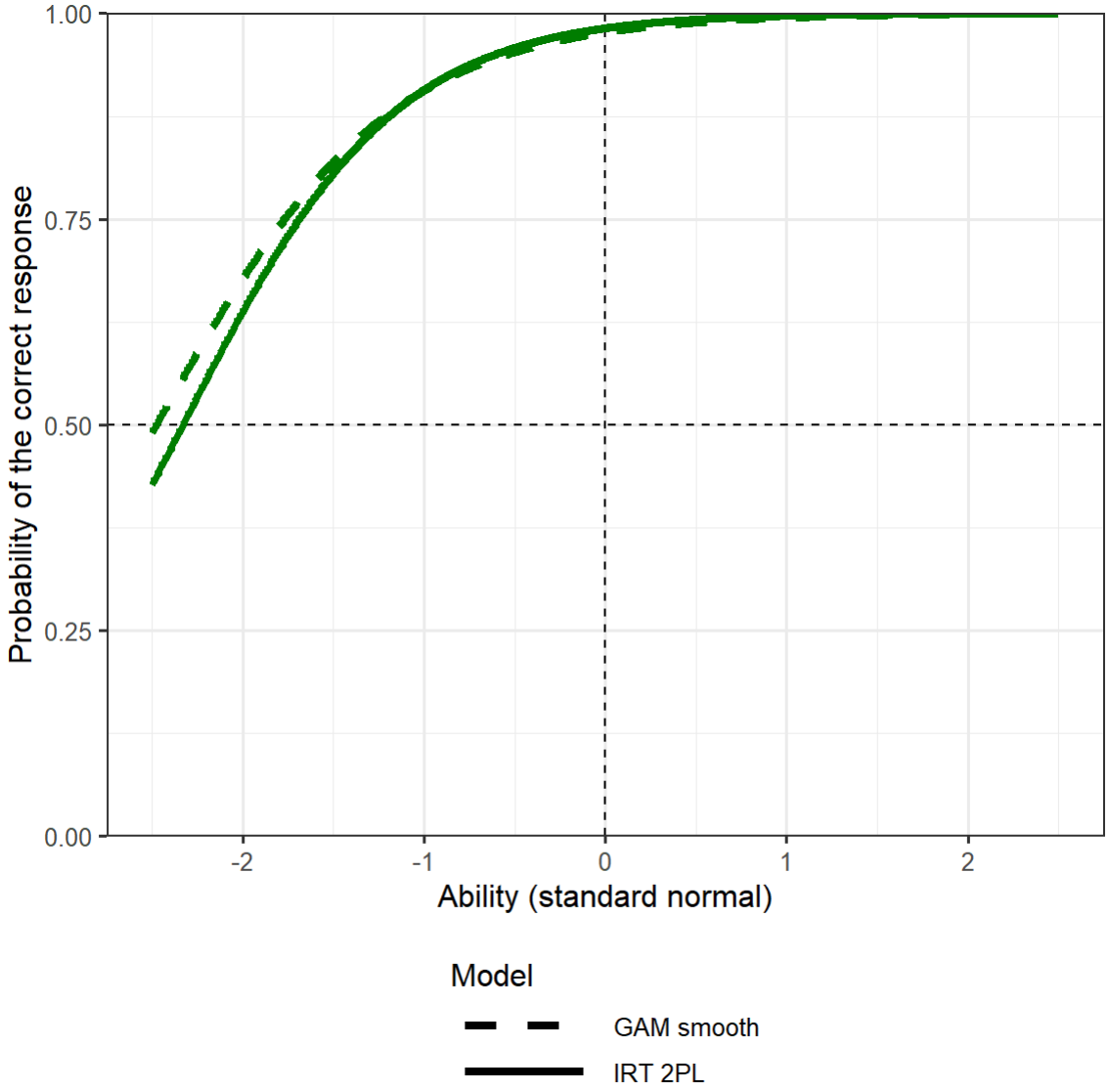
10.2.67.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.0	17.6	36.6	65.8	278.4	3.5	1.0
Middle quartiles	4.2	26.0	40.8	63.9	273.3	3.7	0.7
Upper quartile	13.7	30.9	48.8	75.7	253.5	3.9	0.6
Overall	4.0	25.7	42.5	66.8	278.4	3.7	0.8

10.2.68 S041027 (LLM)

10.2.68.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	94.90%	294	TIMSS	CR	0.30	0.27	1.71	−2.33	0.019

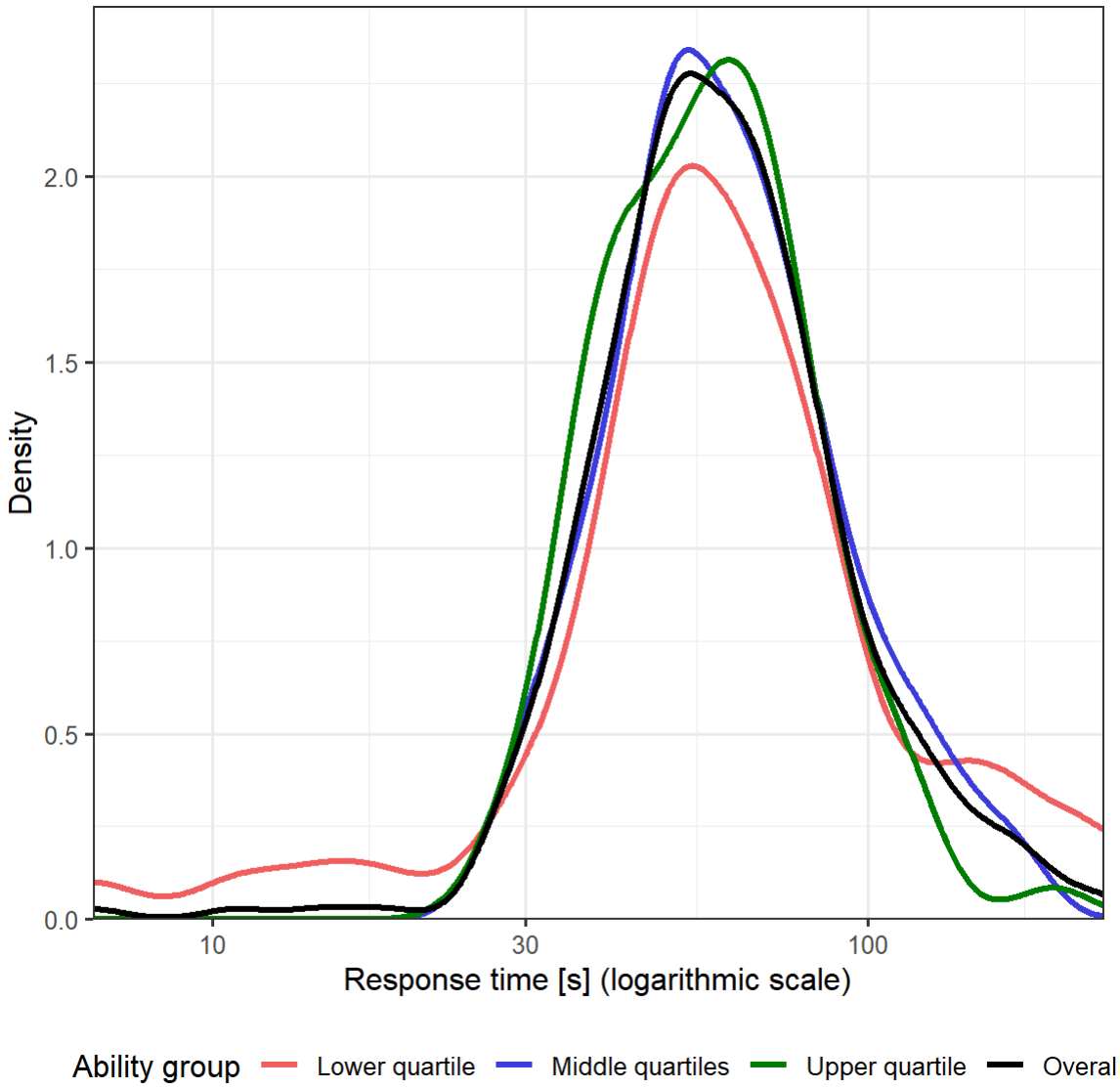
10.2.68.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.67	0.263	1.000	0.023	N/A

10.2.68.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

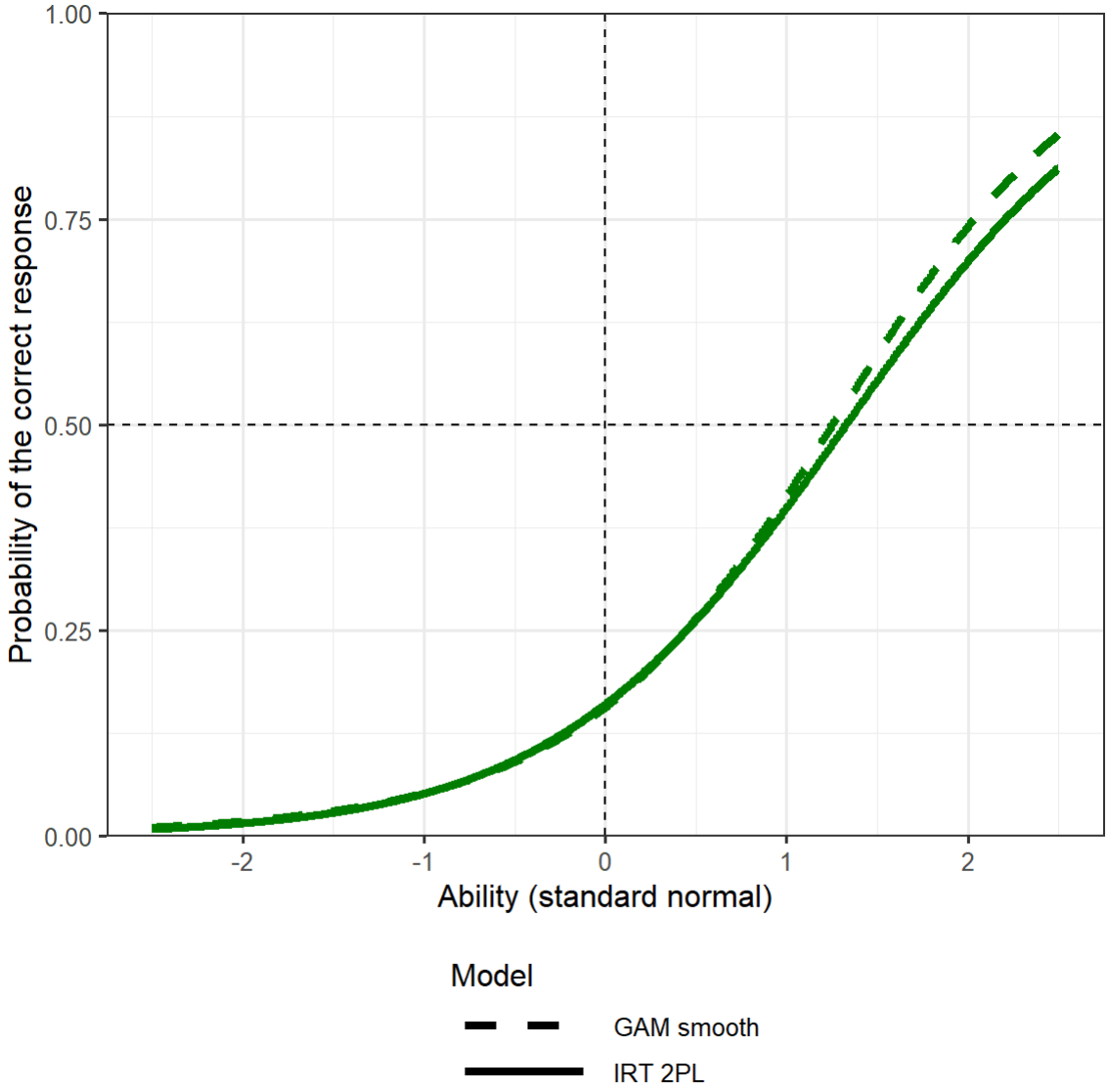
10.2.68.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	6.6	45.6	58.8	79.0	228.1	4.1	0.6
Middle quartiles	27.0	47.4	59.2	78.5	173.5	4.1	0.4
Upper quartile	27.9	43.9	57.0	72.6	191.7	4.0	0.4
Overall	6.6	45.3	58.4	77.3	228.1	4.1	0.4

10.2.69 S061130 (LLM)

10.2.69.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	21.05%	305	TIMSS	CR	0.41	0.32	1.25	1.33	0.024

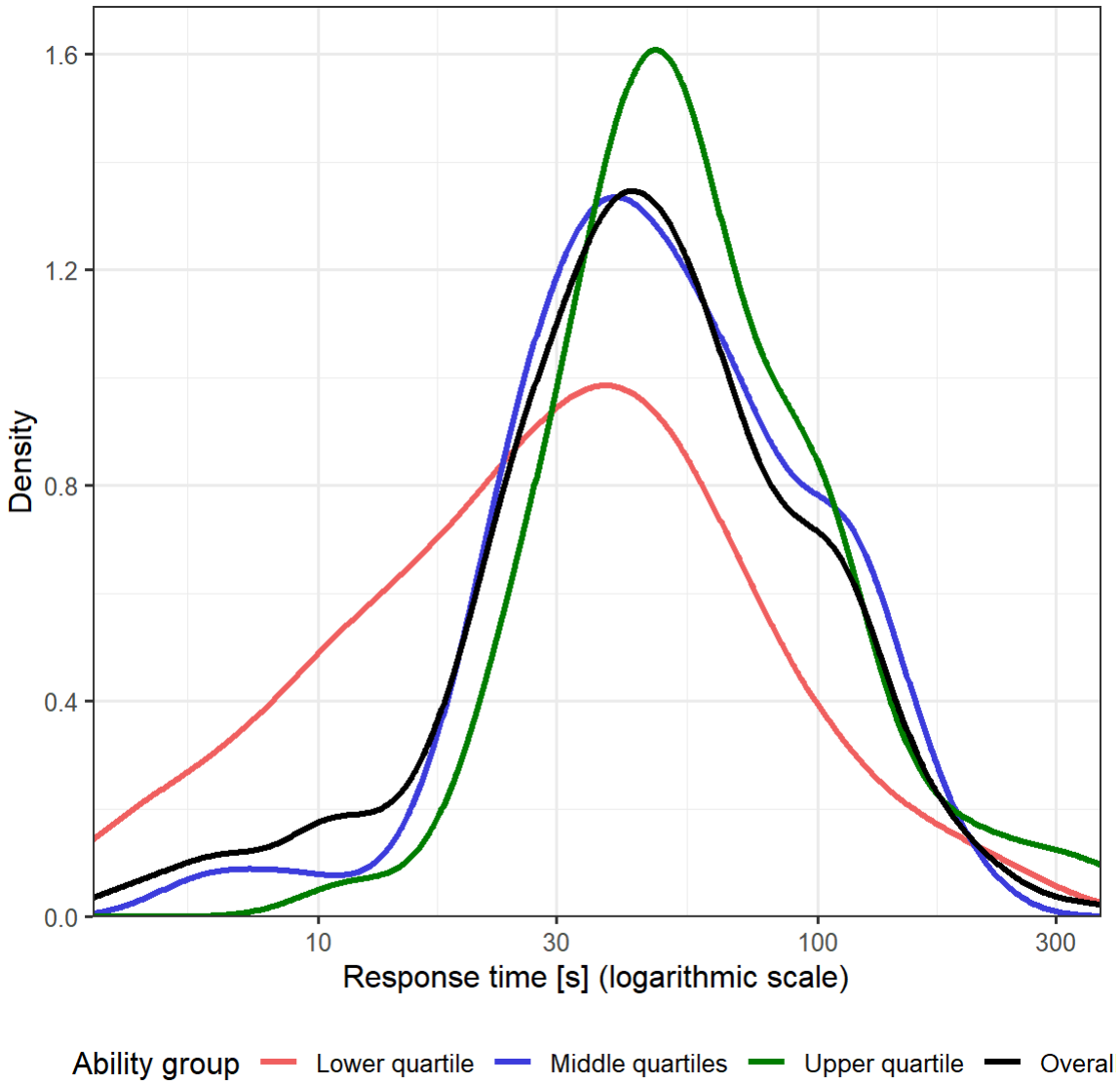
10.2.69.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.64	0.726	1.000	0.002	N/A

10.2.69.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

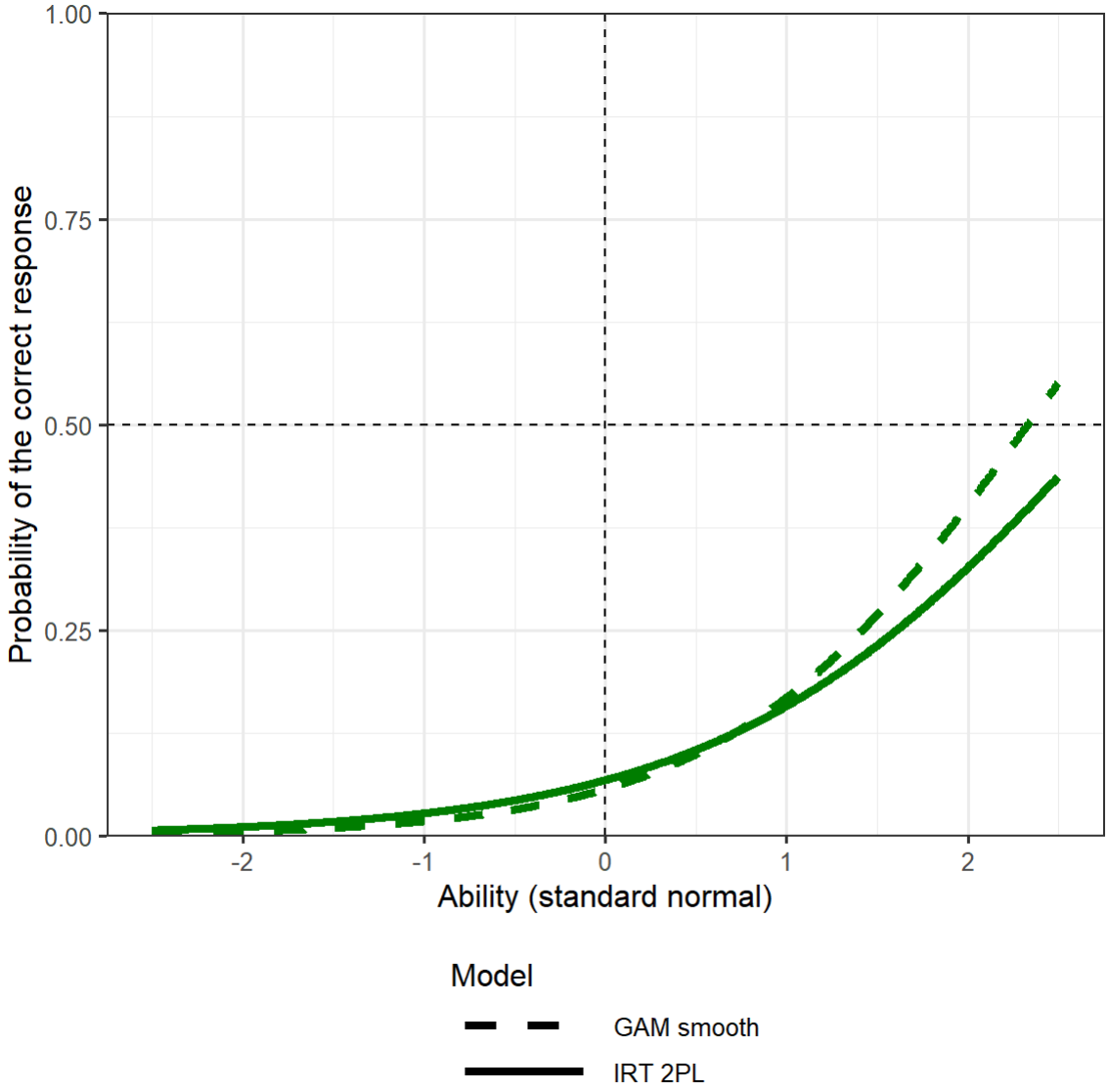
10.2.69.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.5	15.5	30.4	52.3	229.5	3.4	0.9
Middle quartiles	5.2	31.2	46.3	77.6	209.8	3.9	0.7
Upper quartile	11.7	38.8	52.9	89.1	367.8	4.0	0.6
Overall	3.5	28.9	44.1	72.3	367.8	3.8	0.8

10.2.70 S890002 (LLM)

10.2.70.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	9.29%	281	LLM complex setup	CR	0.30	0.13	0.95	2.77	0.00

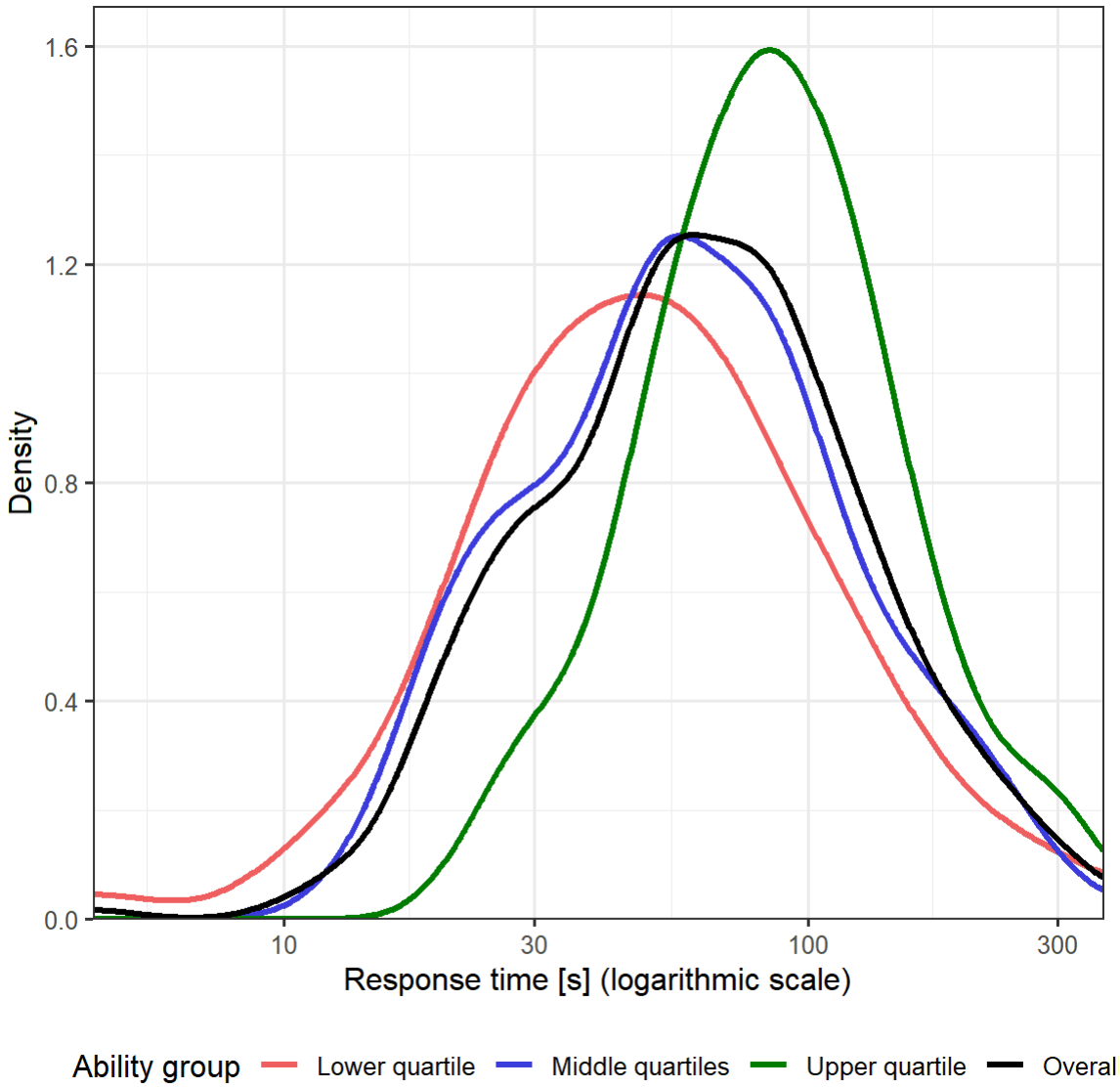
10.2.70.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.68	0.712	1.000	0.004	N/A

10.2.70.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

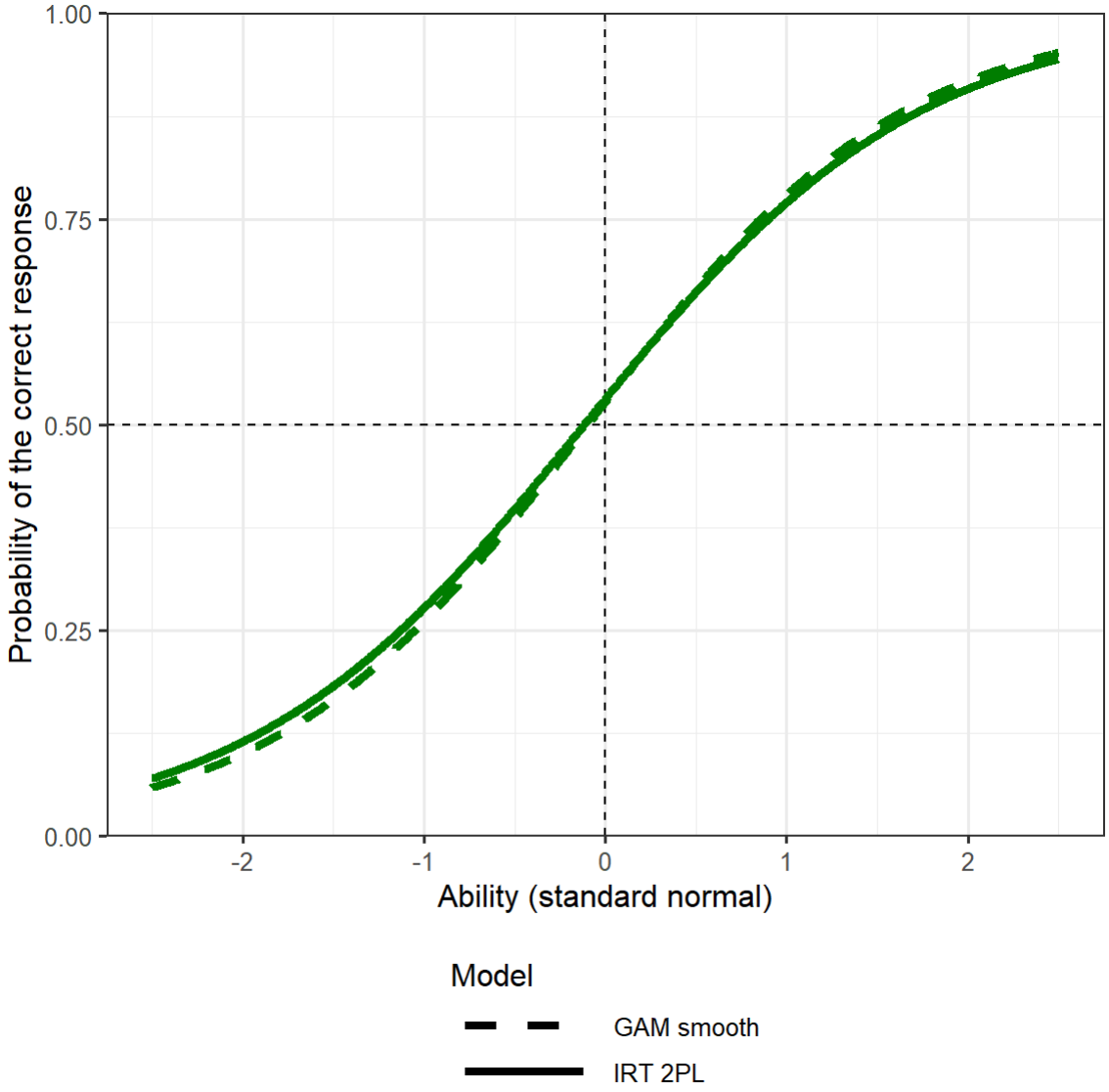
10.2.70.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.3	28.8	50.0	79.9	366.2	3.9	0.8
Middle quartiles	13.1	34.2	56.5	91.5	341.1	4.1	0.7
Upper quartile	24.5	57.0	83.5	124.3	315.7	4.4	0.6
Overall	4.3	38.7	60.6	96.5	366.2	4.1	0.7

10.2.71 S051037 (TIMSS)

10.2.71.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	53.74%	294	TIMSS	CR	0.45	0.44	1.08	−0.11	0.025

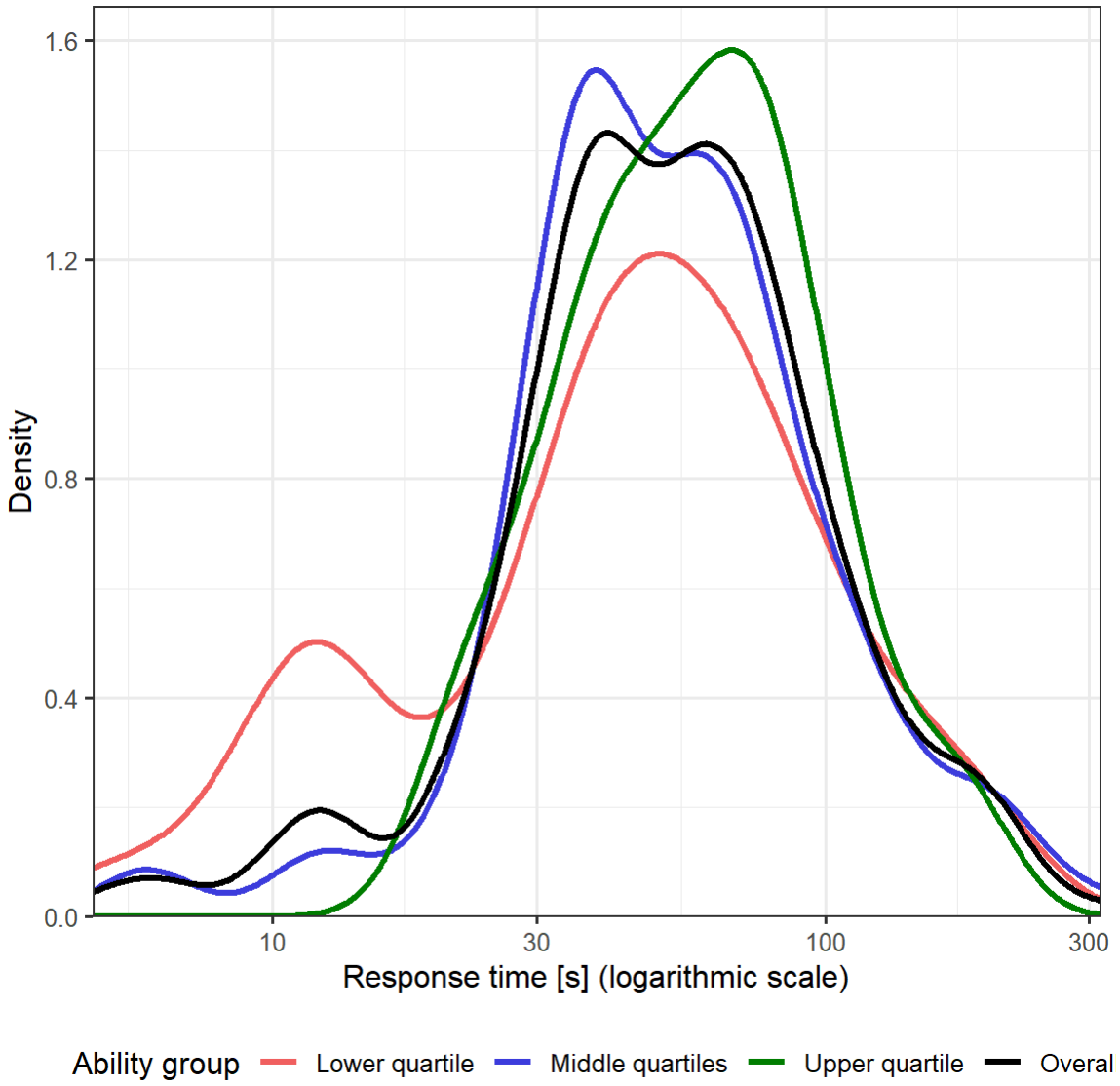
10.2.71.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.54	0.281	1.000	0.007	N/A

10.2.71.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

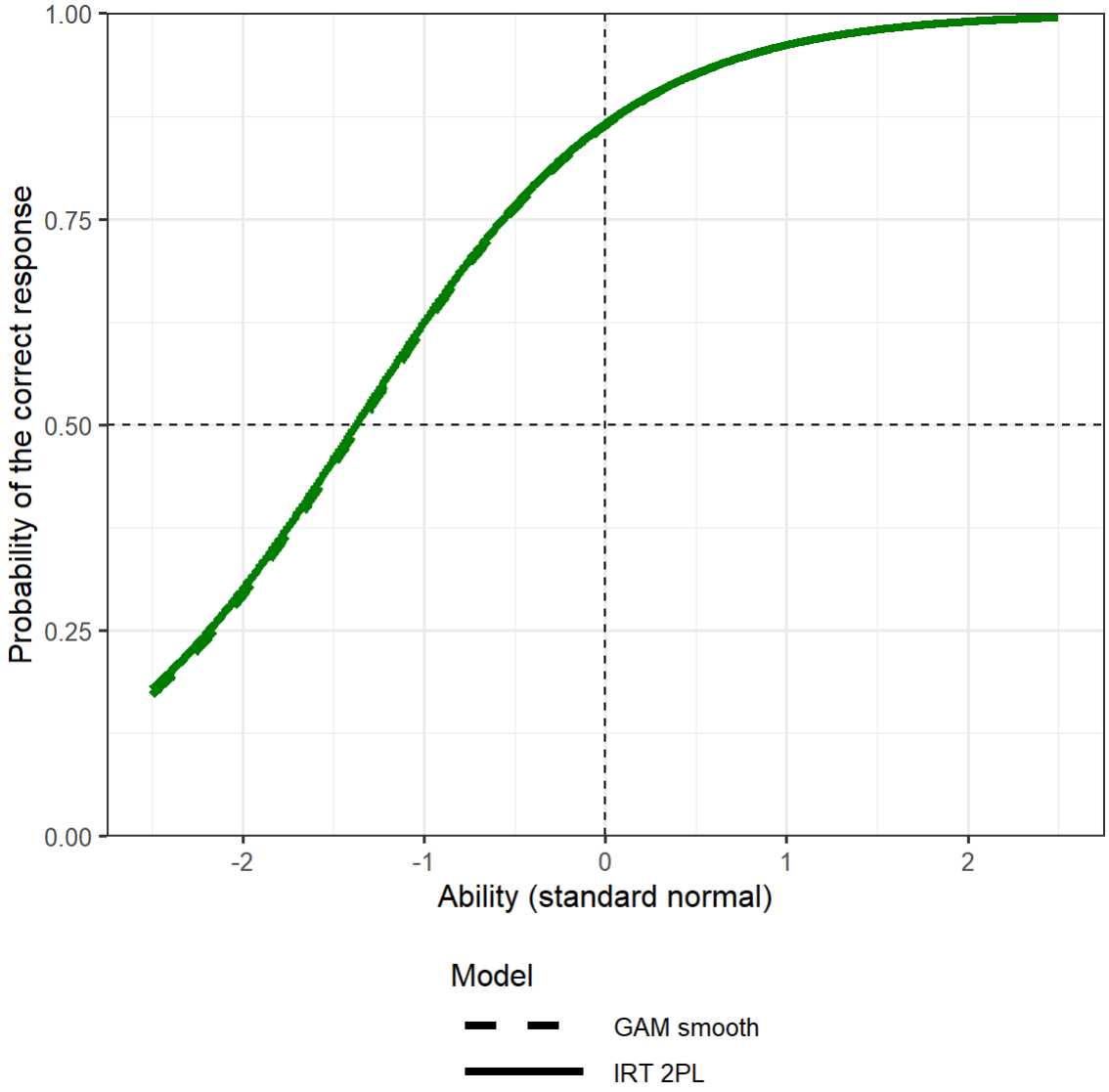
10.2.71.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.7	30.1	45.3	74.7	204.8	3.7	0.9
Middle quartiles	5.3	35.1	51.6	74.5	314.7	3.9	0.7
Upper quartile	19.2	39.1	59.1	80.7	189.9	4.1	0.5
Overall	4.7	35.1	52.3	77.8	314.7	3.9	0.7

10.2.72 S041032 (LLM)

10.2.72.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	81.29%	294	TIMSS	CR	0.43	0.44	1.36	−1.37	0.000

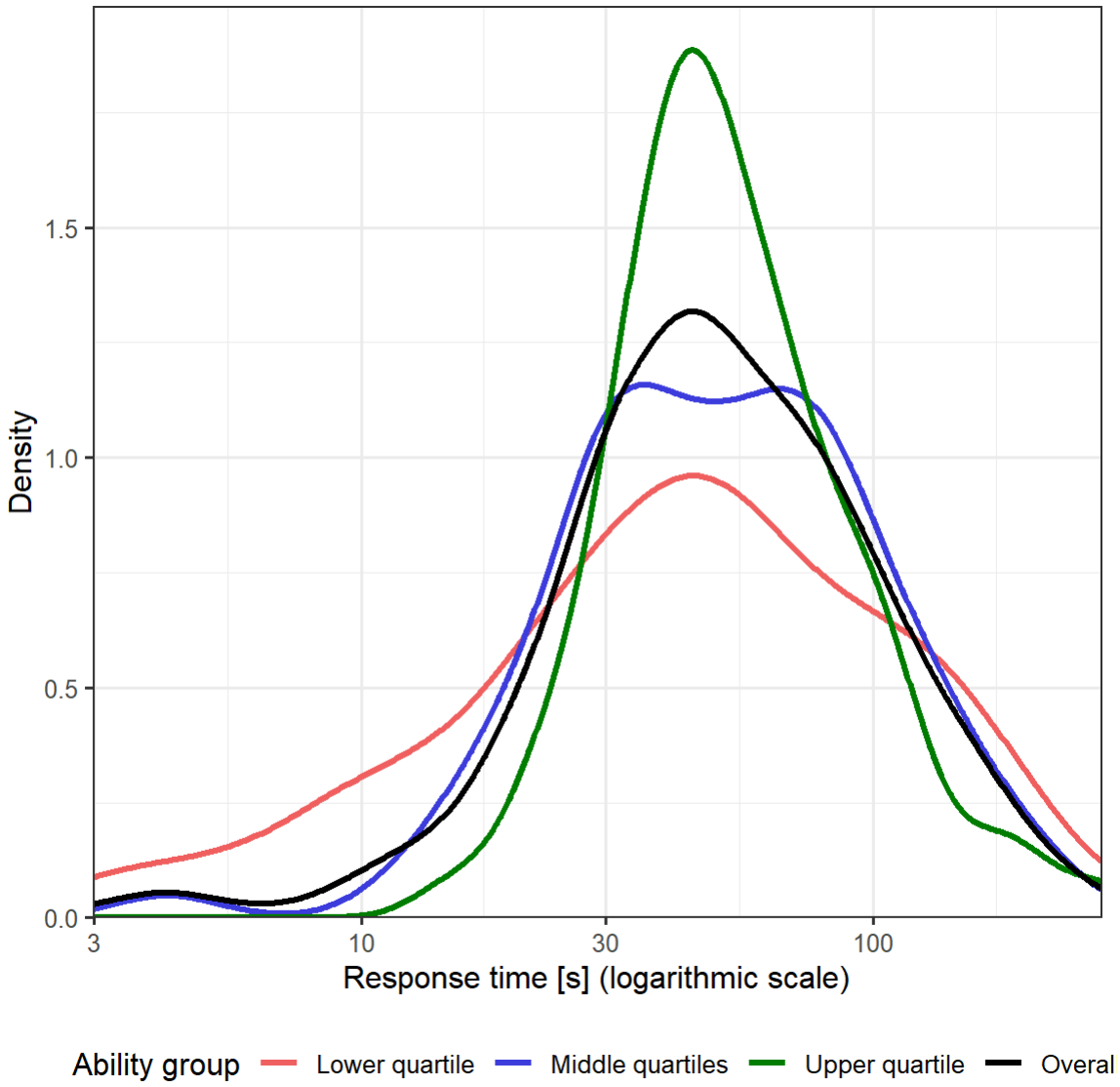
10.2.72.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.94	0.380	1.000	0.008	N/A

10.2.72.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

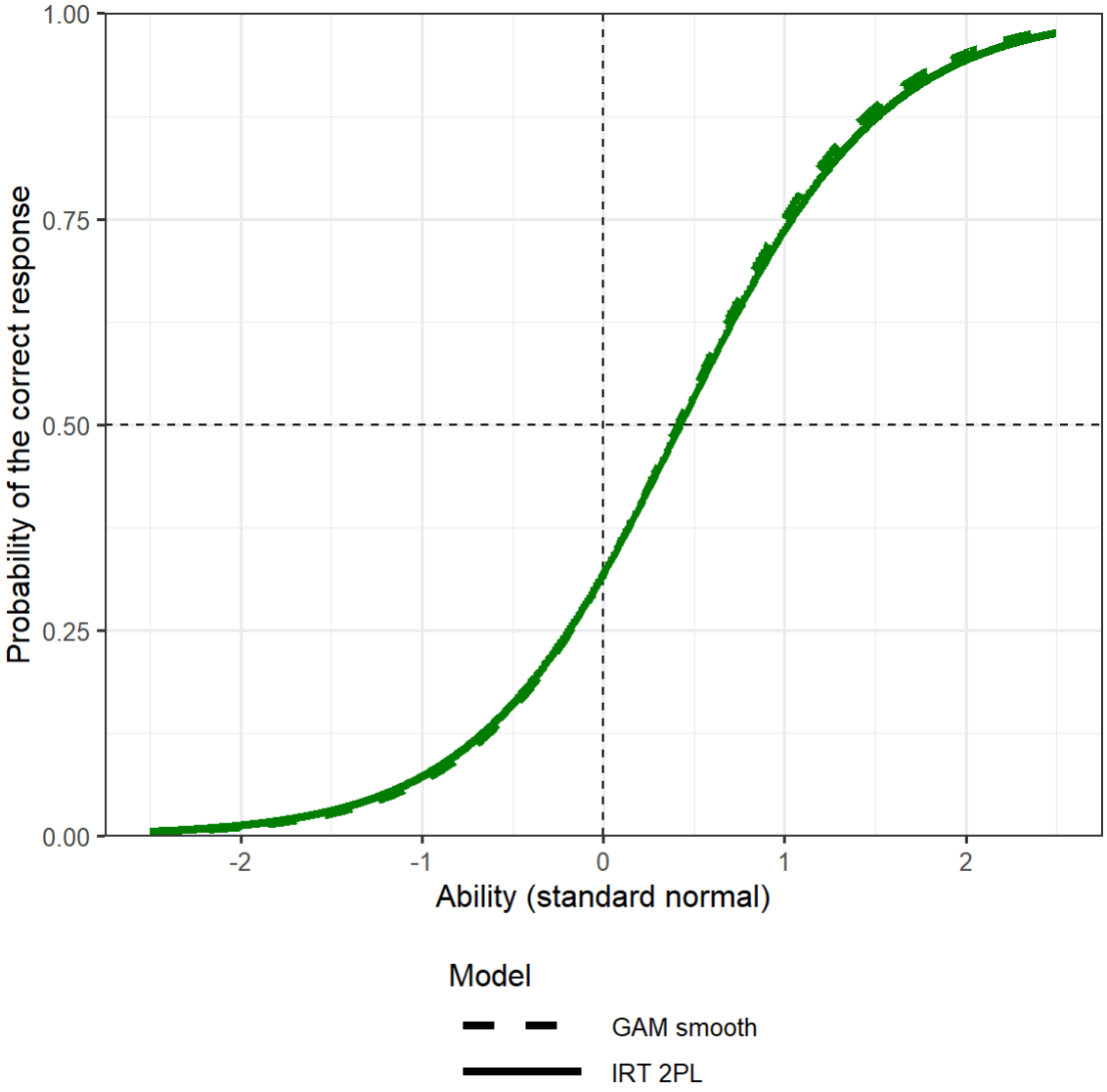
10.2.72.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.0	24.0	45.0	79.5	249.0	3.7	1.0
Middle quartiles	3.9	30.3	50.4	80.6	226.8	3.9	0.7
Upper quartile	15.0	37.0	48.4	69.5	278.3	4.0	0.5
Overall	3.0	31.6	48.0	79.3	278.3	3.9	0.7

10.2.73 SP71077 (LLM)

10.2.73.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	38.93%	299	TIMSS	CR	0.57	0.61	1.79	0.43	0.027

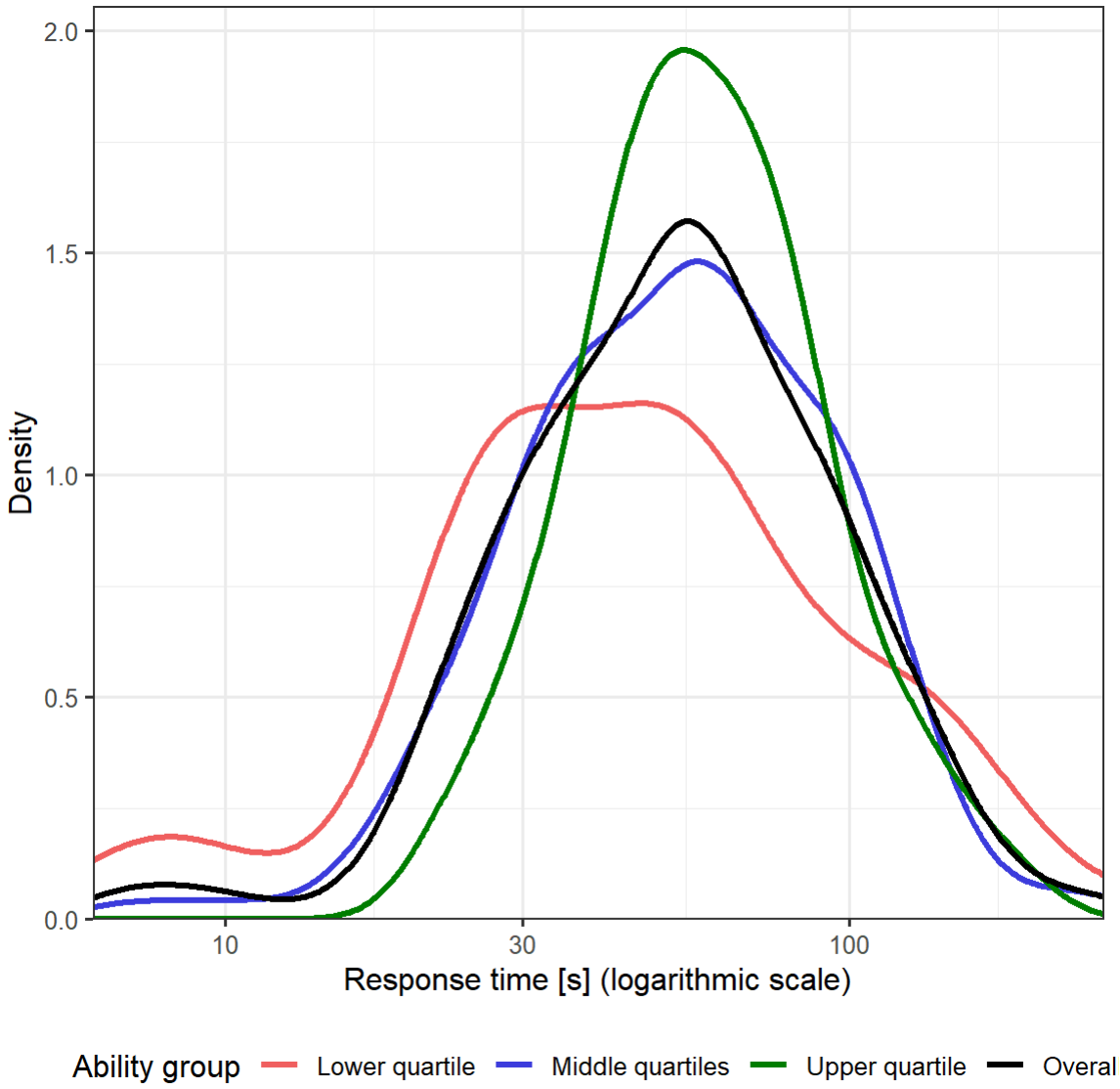
10.2.73.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.08	0.962	1.000	0.000	N/A

10.2.73.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

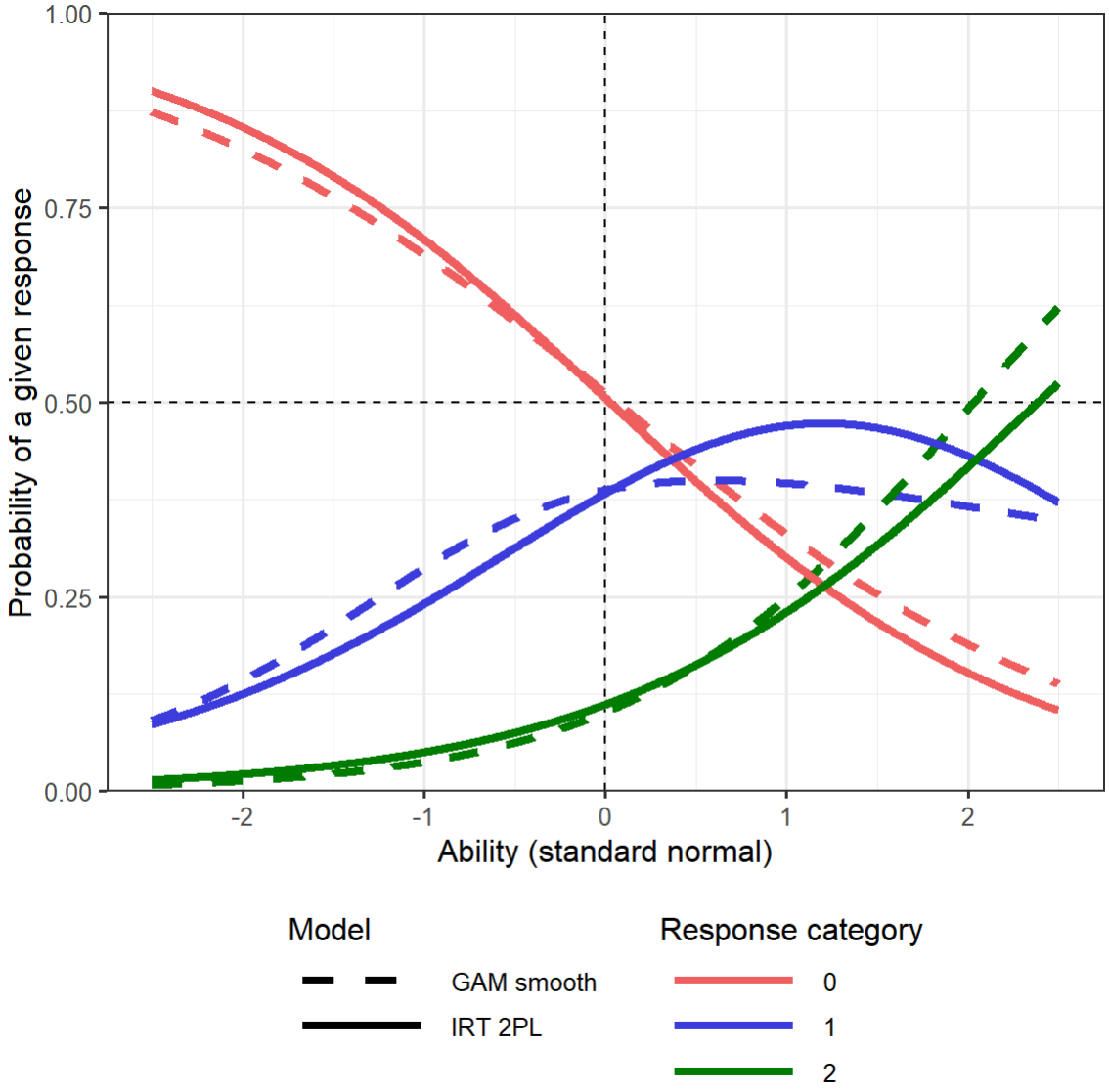
10.2.73.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	6.1	27.2	45.4	71.1	255.5	3.8	0.8
Middle quartiles	7.0	35.2	54.2	79.7	248.0	4.0	0.6
Upper quartile	22.0	42.9	58.8	77.0	187.3	4.1	0.4
Overall	6.1	35.2	53.5	77.9	255.5	3.9	0.6

10.2.74 S041006 (LLM)

10.2.74.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	Catego difficul
2	30.94%	305	TIMSS	CR	0.37	0.34	0.87	1.21	0.0

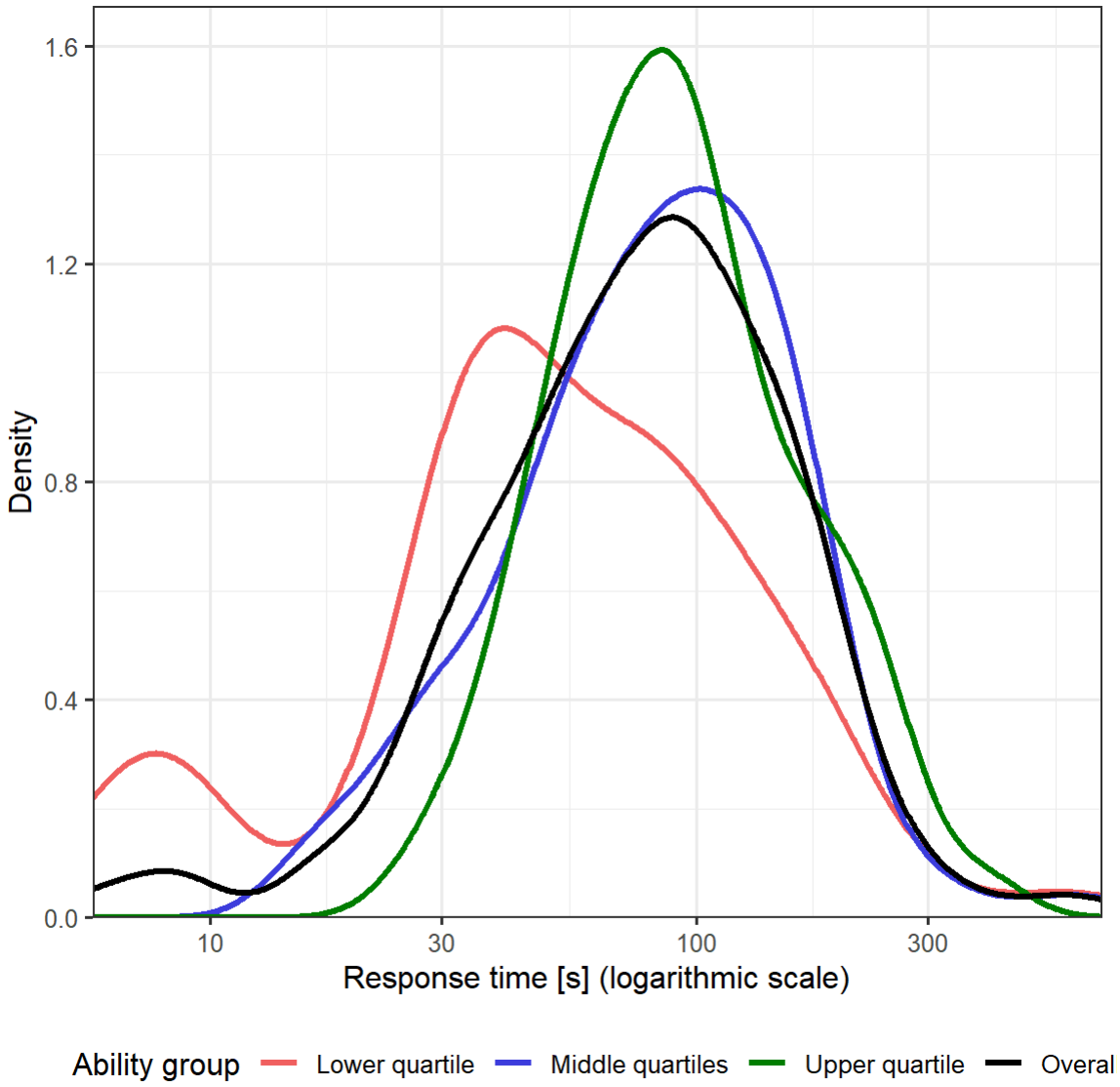
10.2.74.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.49	0.288	1.000	0.007	N/A

10.2.74.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

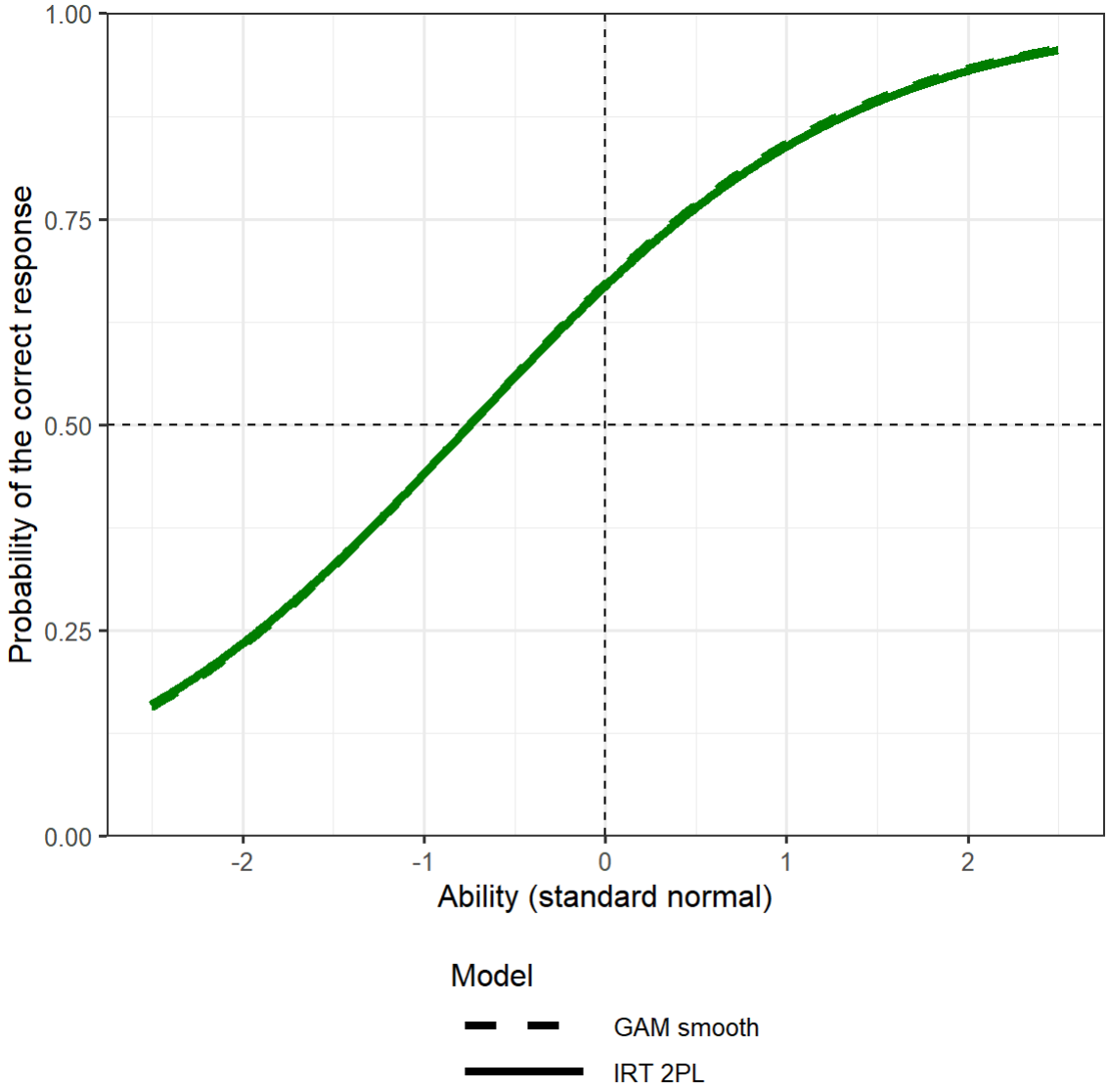
10.2.74.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.7	32.4	52.5	92.6	591.0	3.9	1.0
Middle quartiles	15.0	52.6	84.7	128.0	680.8	4.4	0.7
Upper quartile	26.7	63.0	87.9	125.1	387.5	4.5	0.6
Overall	5.7	49.0	78.7	123.5	680.8	4.3	0.8

10.2.75 S041043 (TIMSS)

10.2.75.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	65.10%	299	TIMSS	CR	0.38	0.42	0.94	−0.74	0.006

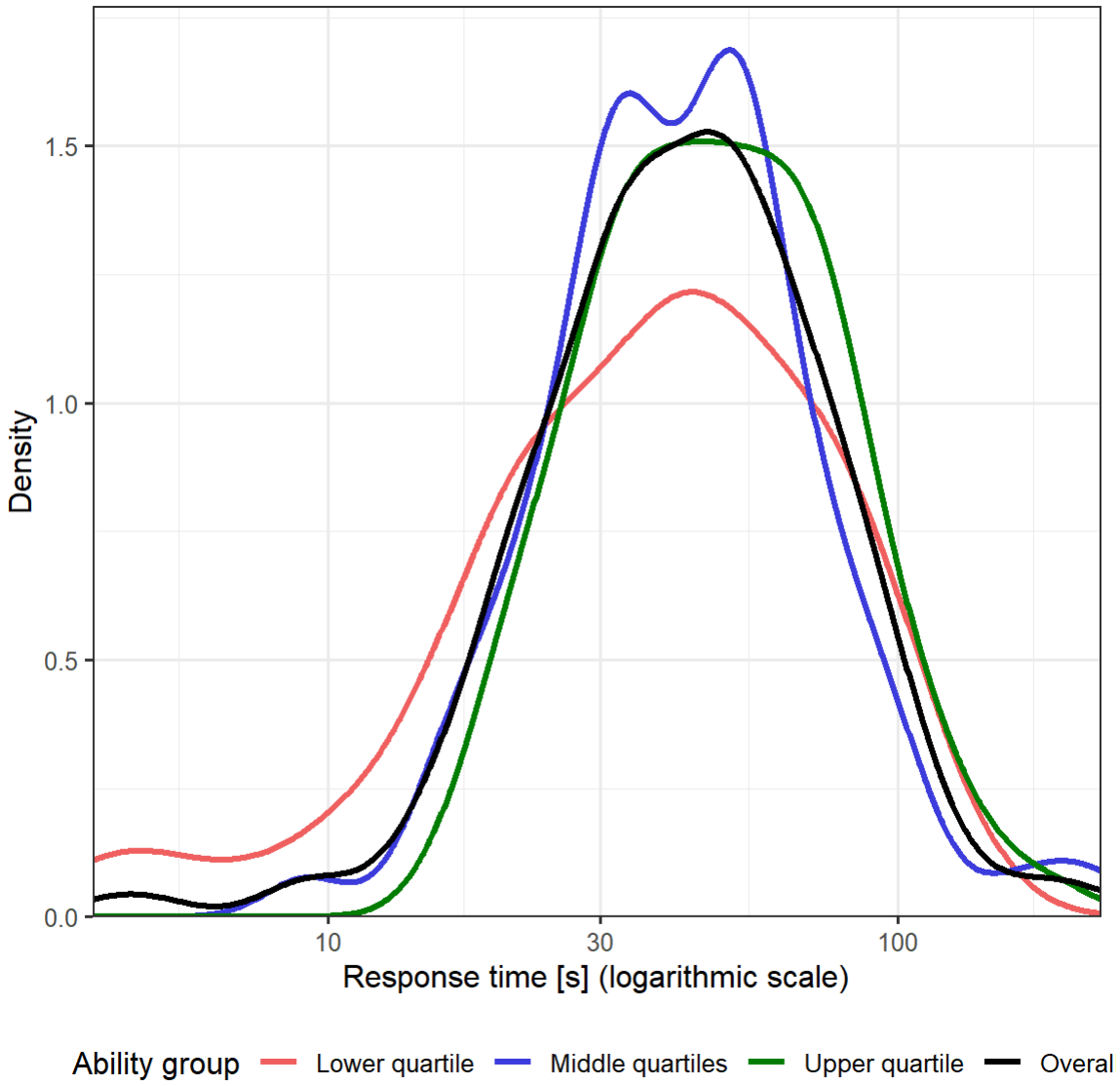
10.2.75.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
2.17	0.337	1.000	0.006	N/A

10.2.75.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

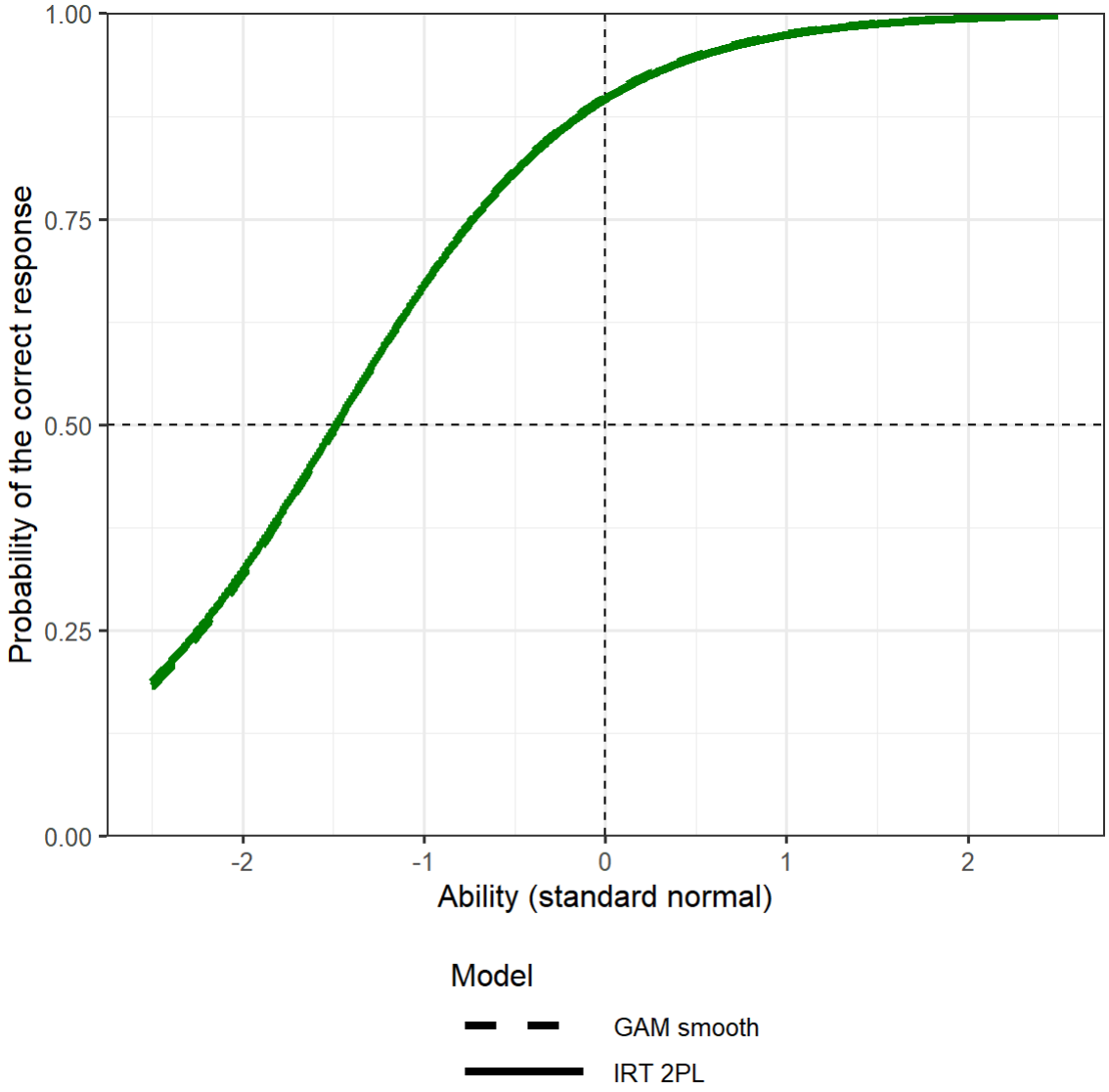
10.2.75.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.9	22.4	40.1	64.0	118.4	3.6	0.7
Middle quartiles	8.8	29.8	42.1	56.6	226.6	3.7	0.6
Upper quartile	17.2	32.9	47.1	67.3	184.4	3.9	0.5
Overall	3.9	28.7	42.9	63.3	226.6	3.7	0.6

10.2.76 S090001 (TIMSS)

10.2.76.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMS generated items Discrimination	Item difficulty	RMSE
1	83.56%	299	LLM teachers' selection	CR	0.43	0.25	1.45	−1.49	0.0(

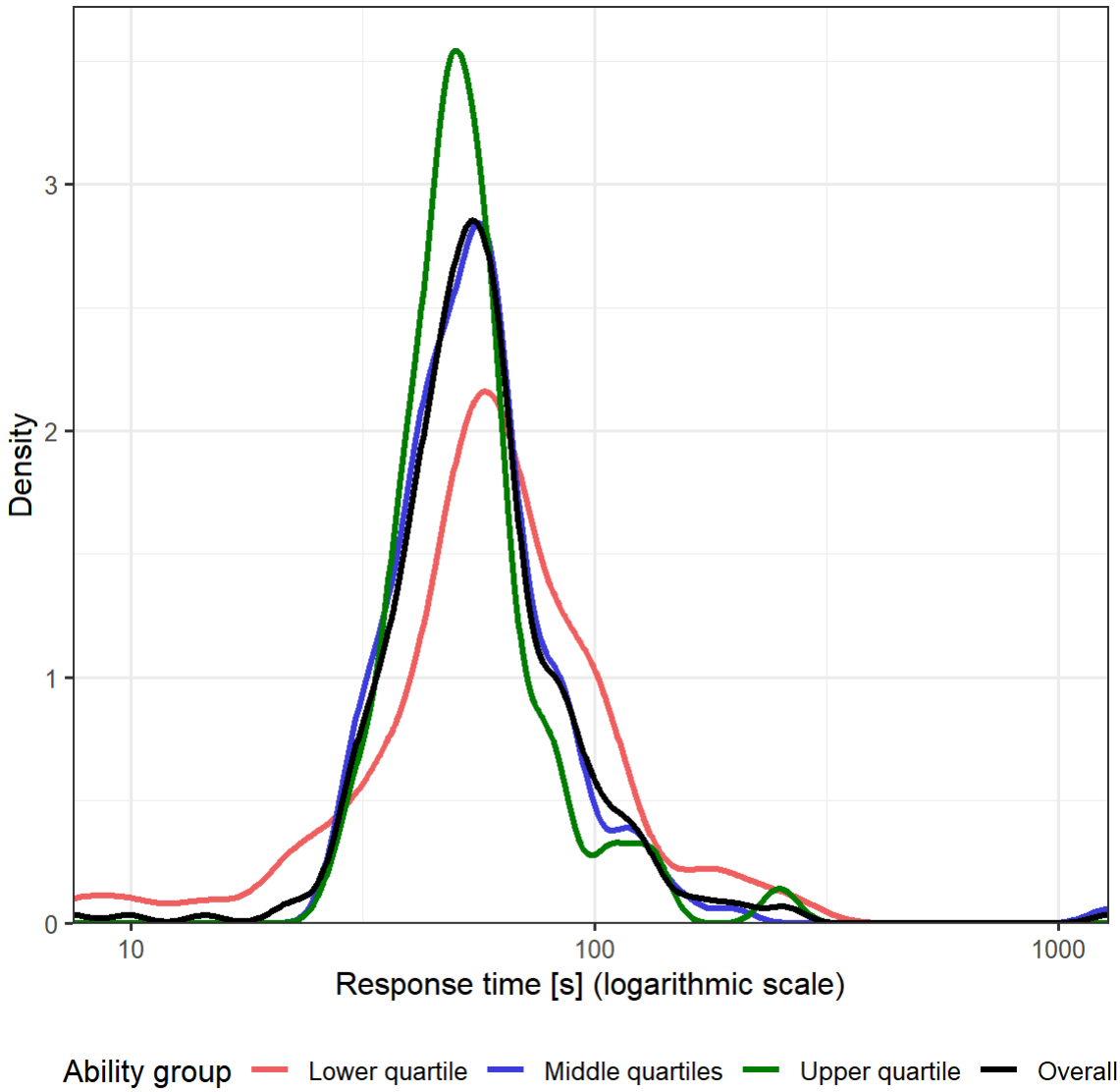
10.2.76.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.04	0.978	1.000	0.000	N/A

10.2.76.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

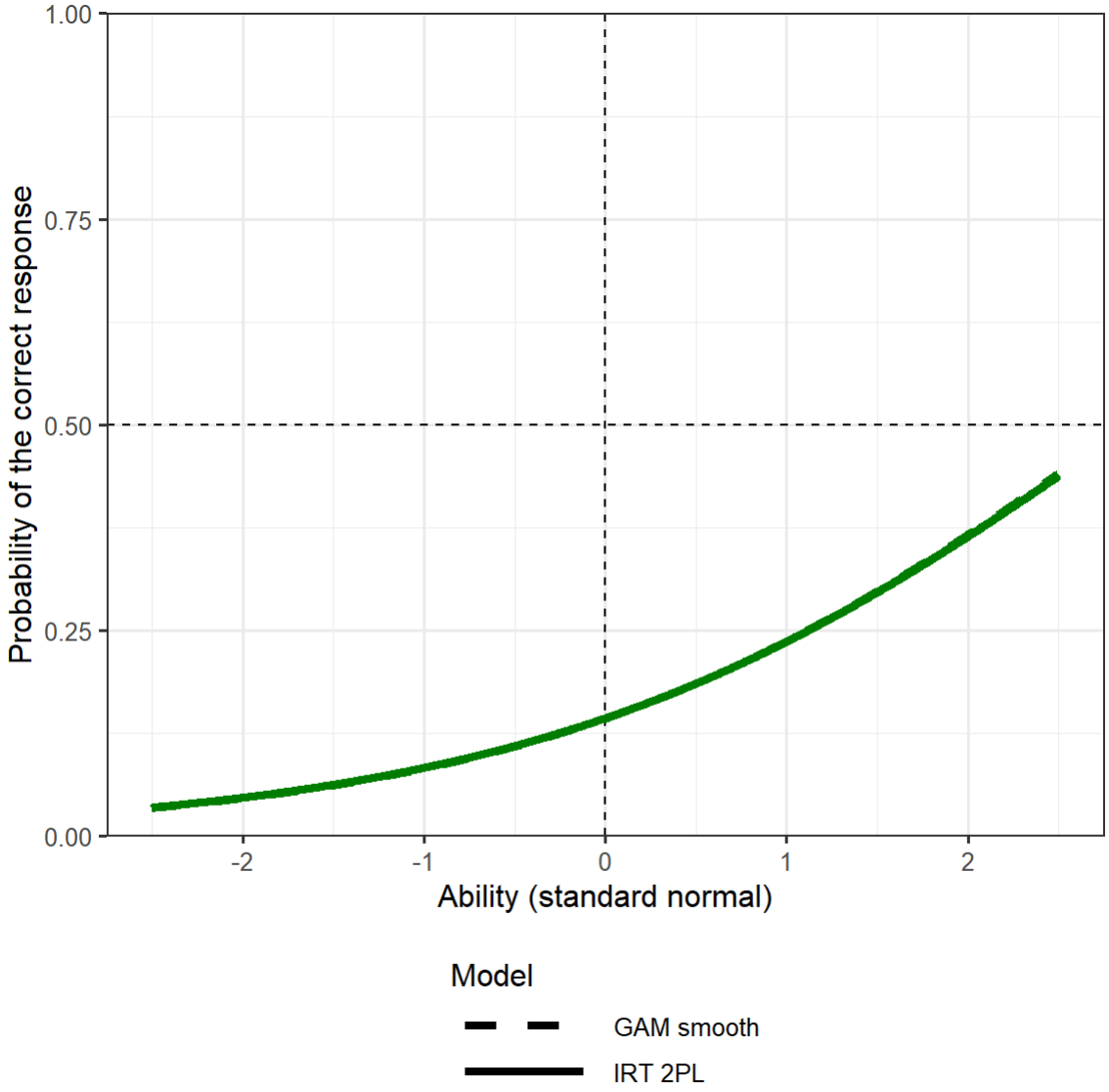
10.2.76.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	7.5	47.1	60.5	81.4	265.9	4.1	0.6
Middle quartiles	28.7	43.4	54.4	65.9	1,279.2	4.0	0.4
Upper quartile	28.0	43.5	49.8	60.1	250.9	4.0	0.4
Overall	7.5	43.7	54.1	66.6	1,279.2	4.0	0.5

10.2.77 S100024 (TIMSS)

10.2.77.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS a generated items Discrimination	Item difficulty	RMSEA
1	15.79%	305	LLM simple setup	CR	0.21	0.11	0.61	2.91	0.000

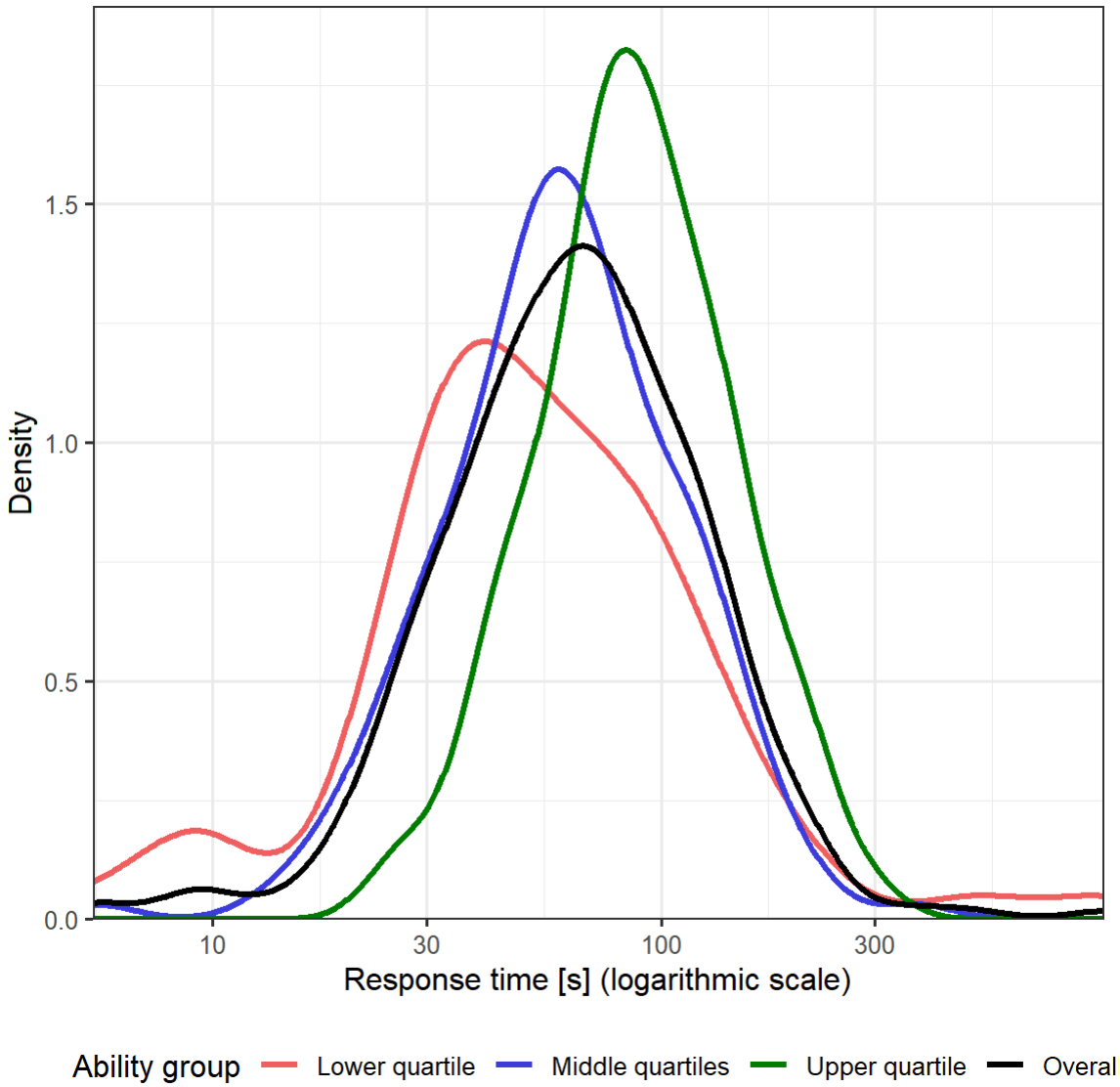
10.2.77.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
5.70	0.058	1.000	0.023	N/A

10.2.77.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

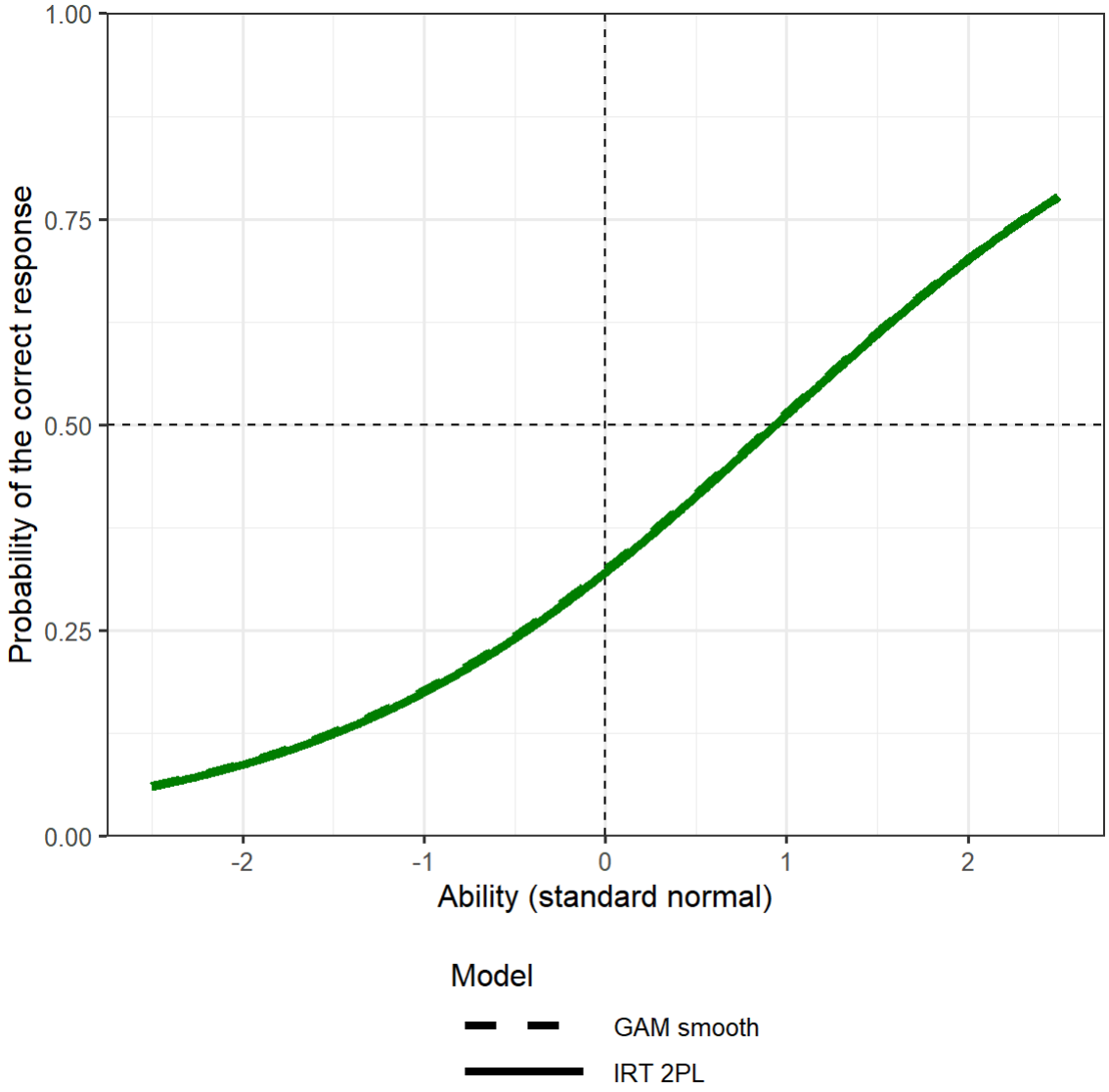
10.2.77.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.4	31.9	48.5	85.3	968.7	3.9	0.9
Middle quartiles	5.6	40.4	60.3	88.3	375.1	4.1	0.6
Upper quartile	25.5	65.5	85.7	123.4	288.2	4.5	0.5
Overall	5.4	42.3	64.8	99.1	968.7	4.1	0.7

10.2.78 S061060 (LLM)

10.2.78.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	34.56%	299	TIMSS	CR	0.33	0.29	0.80	0.94	0.014

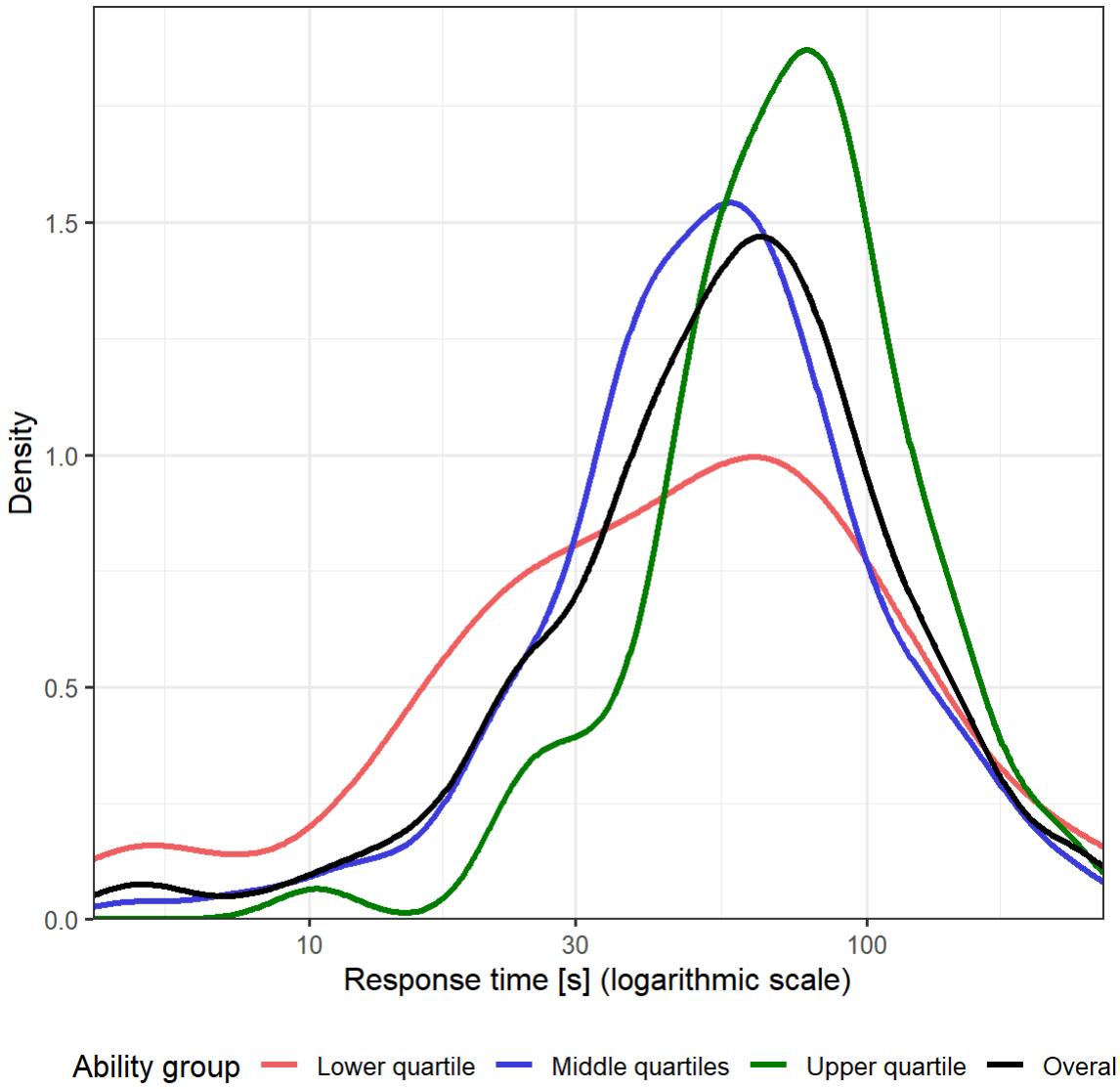
10.2.78.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.42	0.492	1.000	0.004	N/A

10.2.78.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

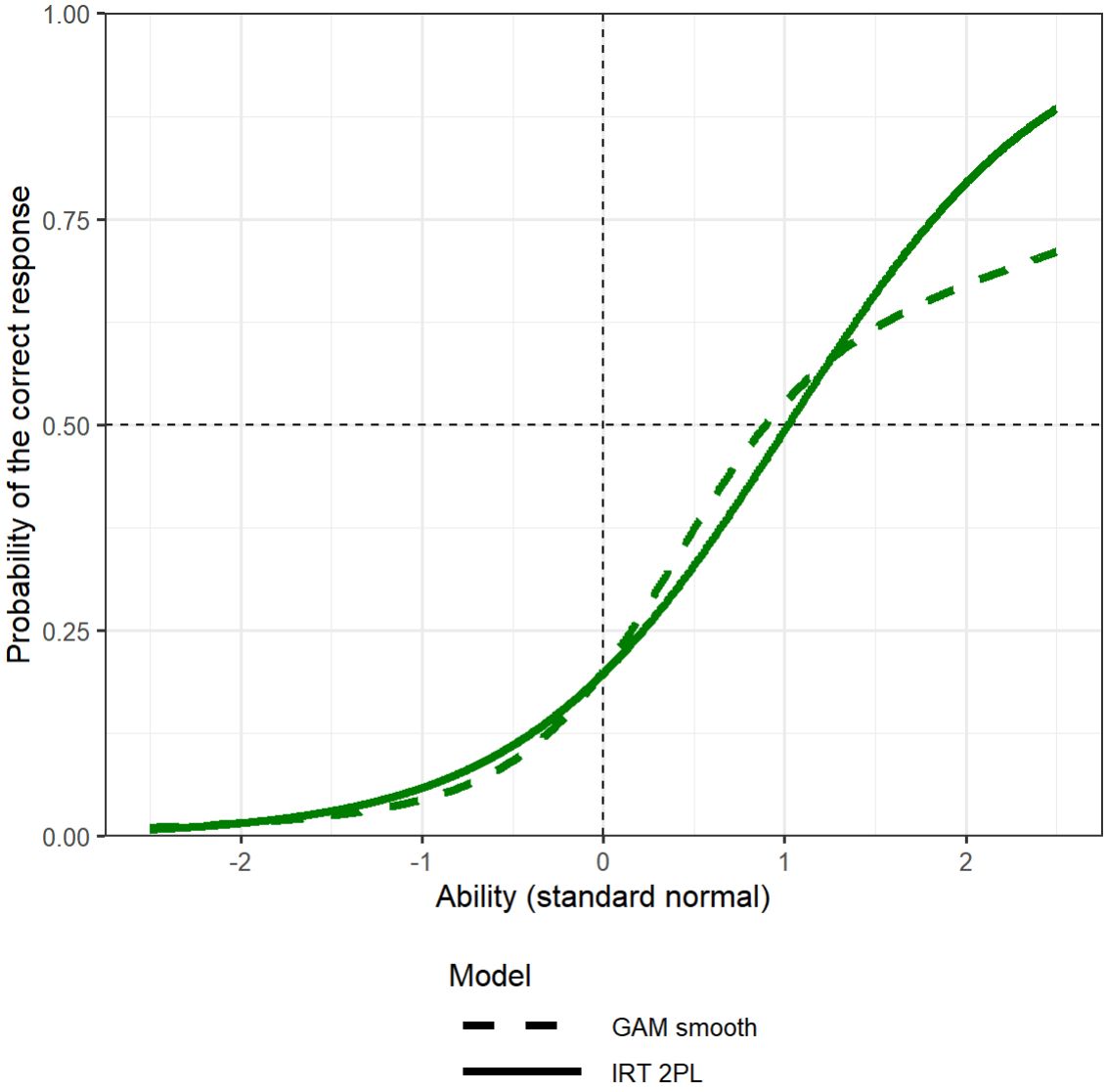
10.2.78.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.1	24.5	48.0	74.3	265.6	3.8	0.9
Middle quartiles	4.7	36.7	55.2	78.3	253.2	4.0	0.7
Upper quartile	10.3	52.0	73.9	98.0	231.6	4.3	0.5
Overall	4.1	37.3	58.1	85.1	265.6	4.0	0.7

10.2.79 S051147 (LLM)

10.2.79.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	27.21%	294	TIMSS	CR	0.44	0.47	1.37	1.02	0.026

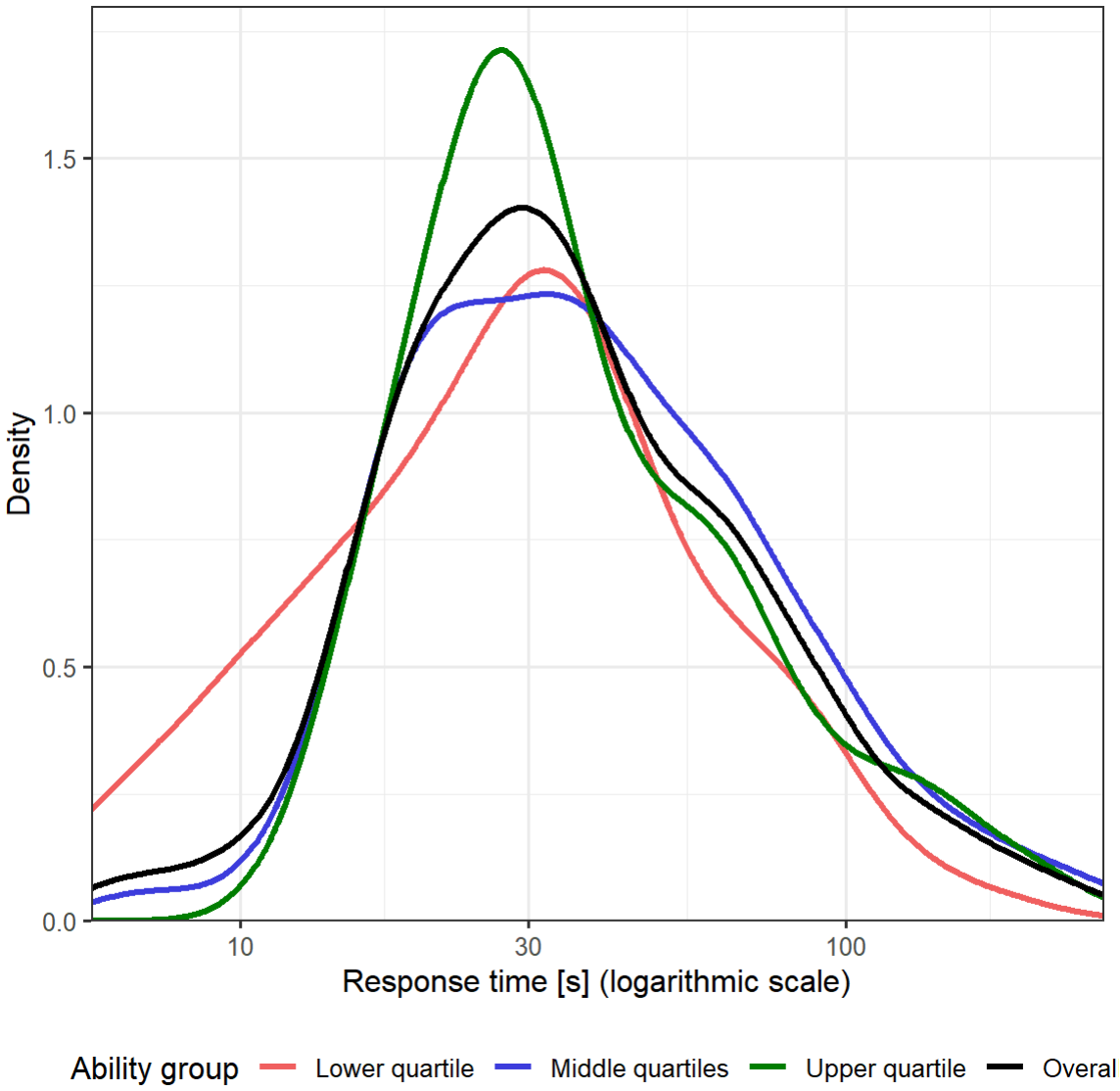
10.2.79.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
4.29	0.117	1.000	0.014	N/A

10.2.79.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

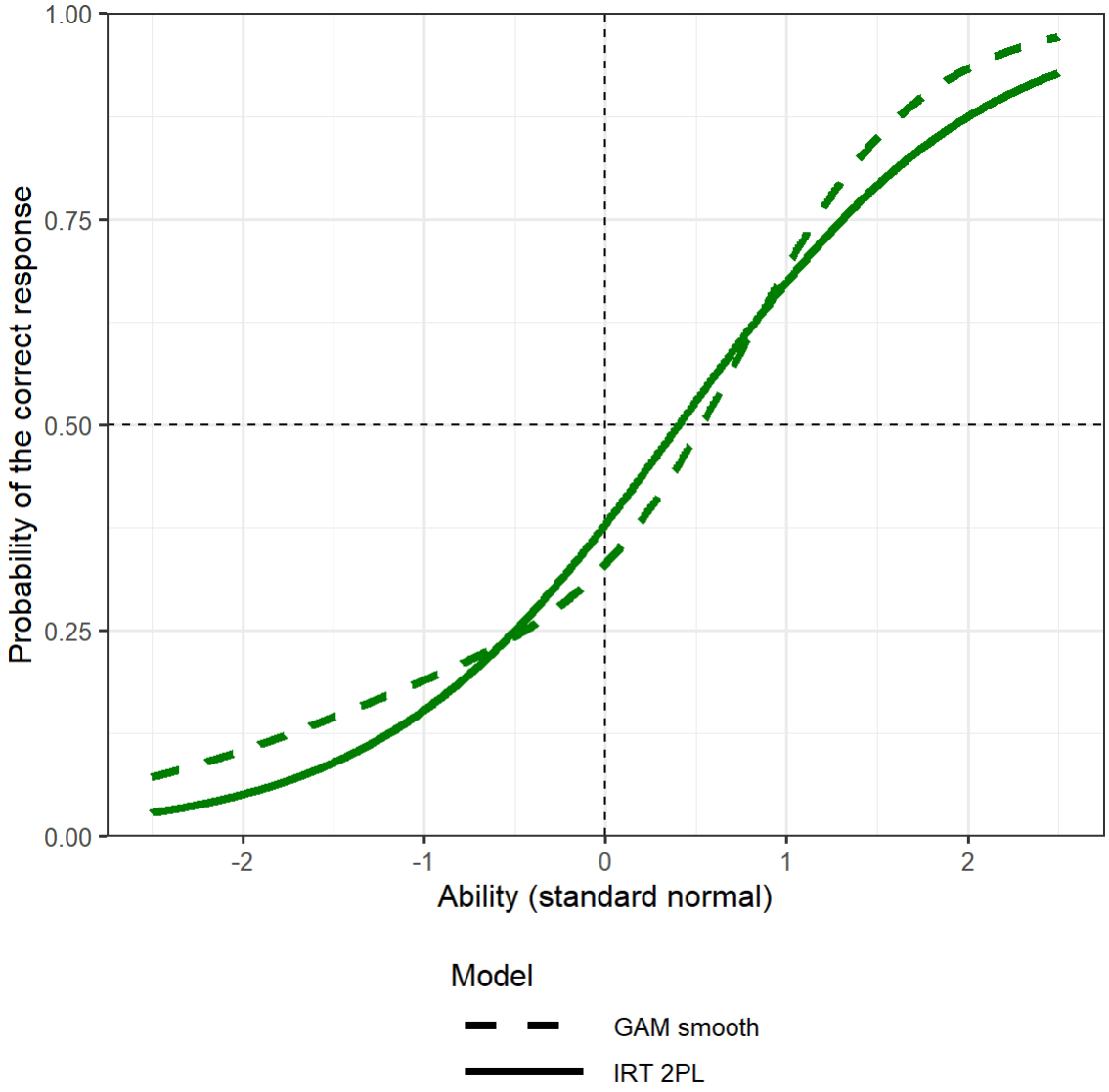
10.2.79.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.7	15.8	29.9	41.5	157.1	3.3	0.7
Middle quartiles	6.7	21.8	35.0	60.5	266.4	3.6	0.7
Upper quartile	14.2	24.0	29.7	56.7	213.6	3.6	0.6
Overall	5.7	21.5	32.3	56.8	266.4	3.6	0.7

10.2.80 S061031 (LLM)

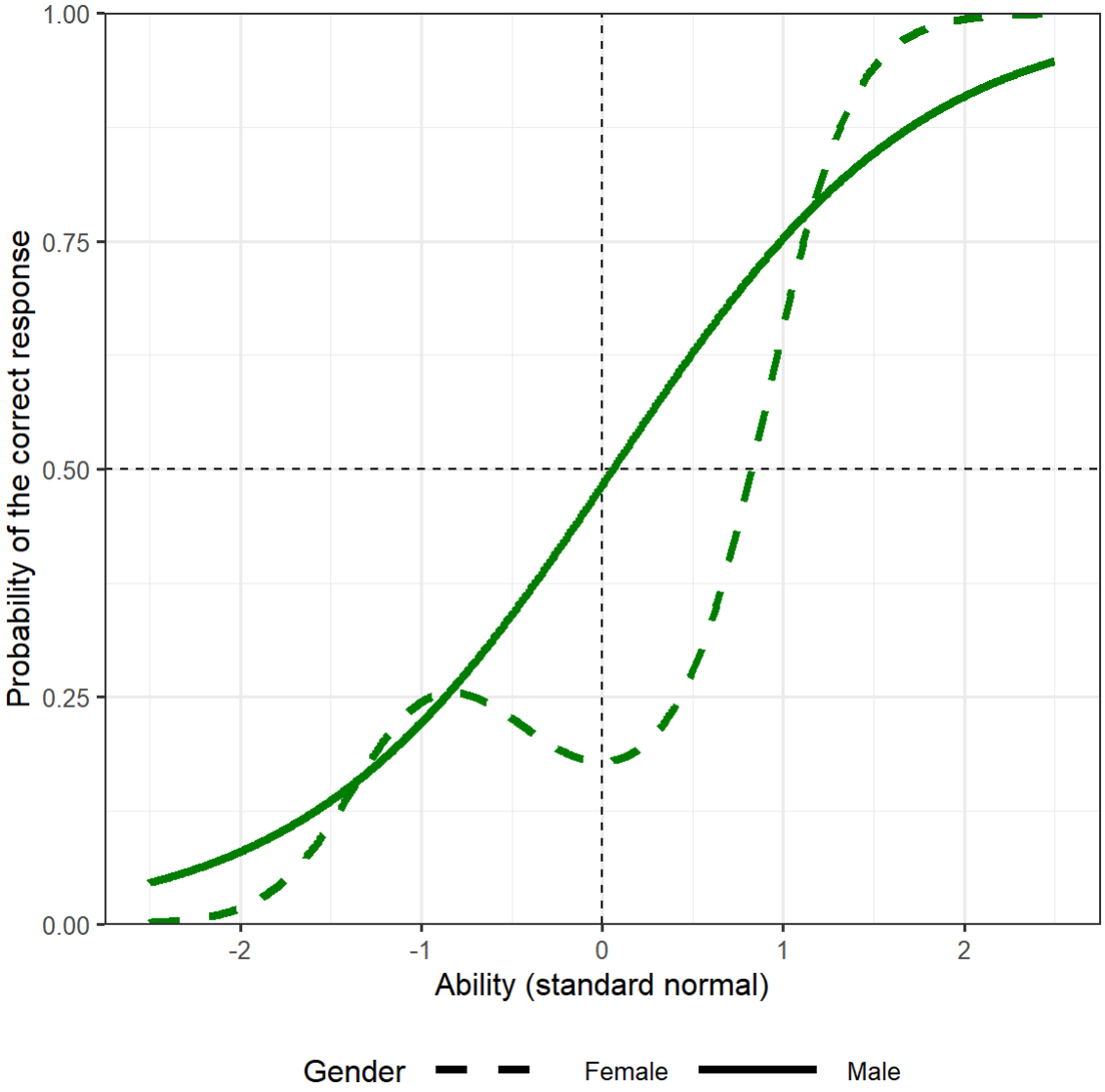
10.2.80.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	41.84%	294	TIMSS	CR	0.45	0.44	1.22	0.41	0.028

10.2.80.2 Gender DIF:

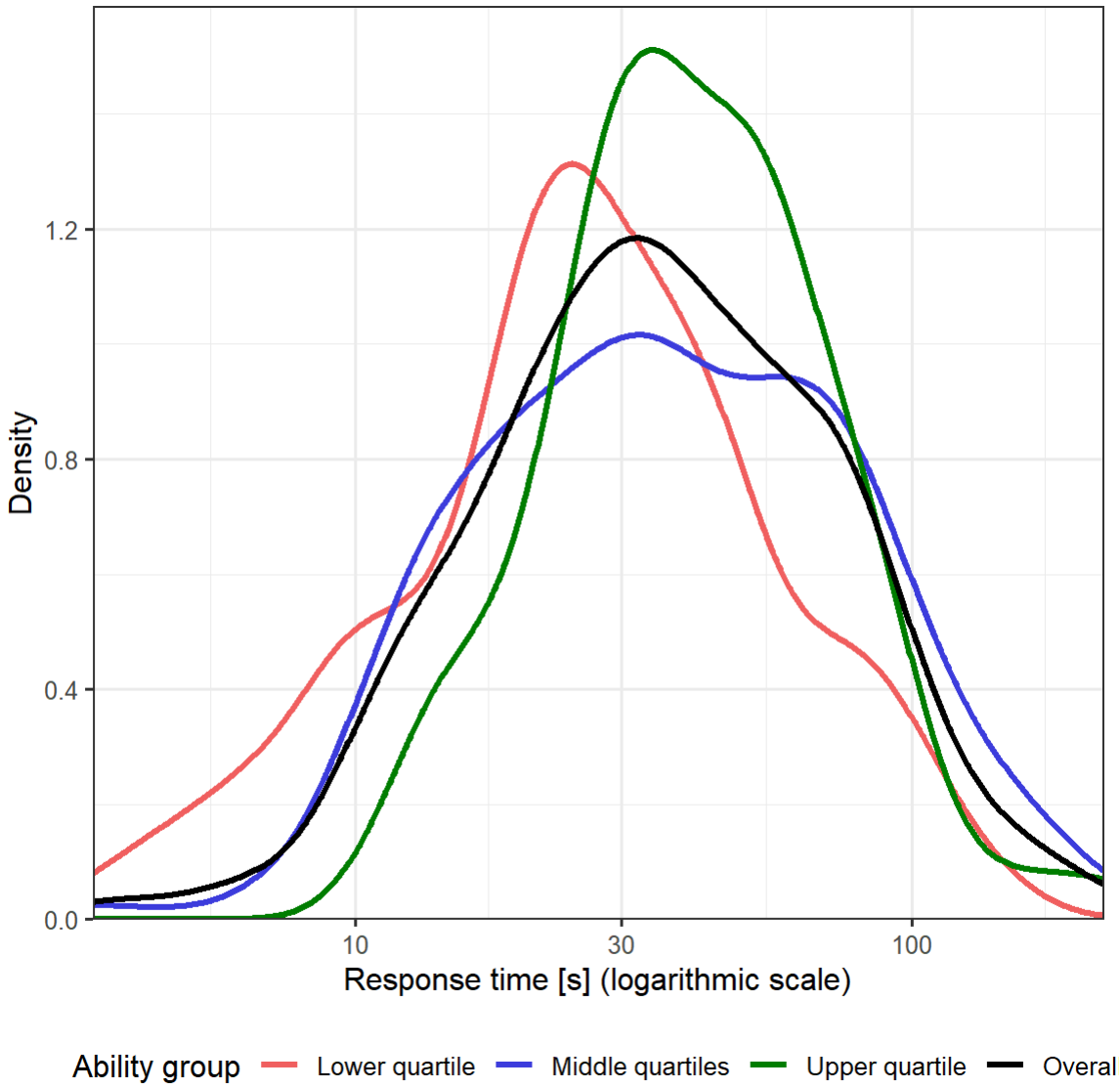
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
17.06	0.000	0.018	0.049	moderate



10.2.80.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

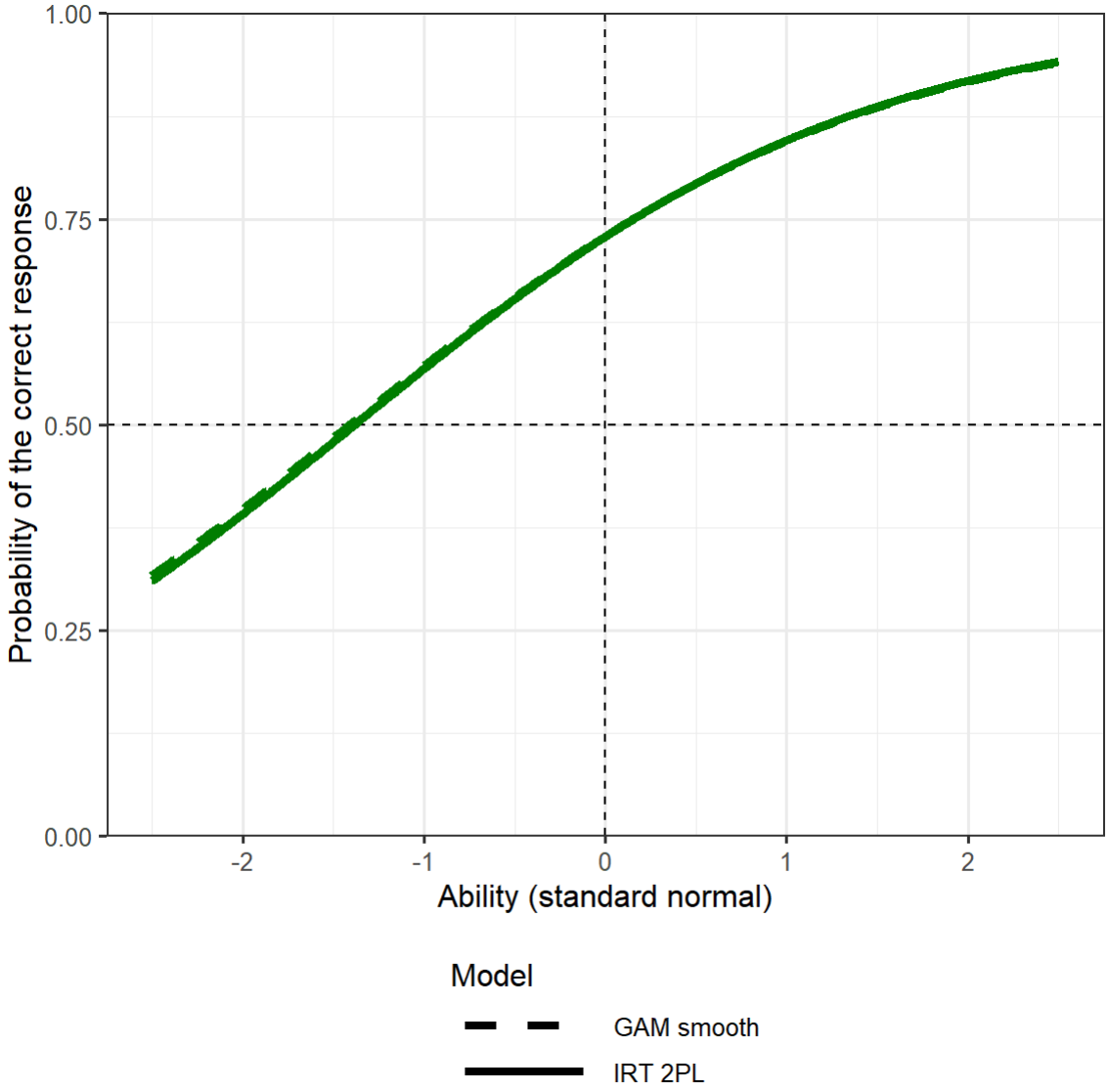
10.2.80.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.8	18.6	25.5	43.3	126.8	3.2	0.8
Middle quartiles	3.4	20.9	35.4	66.1	209.3	3.6	0.8
Upper quartile	11.7	27.9	37.7	58.4	220.6	3.7	0.6
Overall	3.4	21.5	33.9	59.3	220.6	3.5	0.7

10.2.81 S890004 (LLM)

10.2.81.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS generated items: Discrimination	IRT model including both TIMSS generated items: Item difficulty	IRT model including both TIMSS generated items: RMSE
1	70.71%	281	LLM complex setup	CR	0.30	0.23	0.71	−1.38	0.01

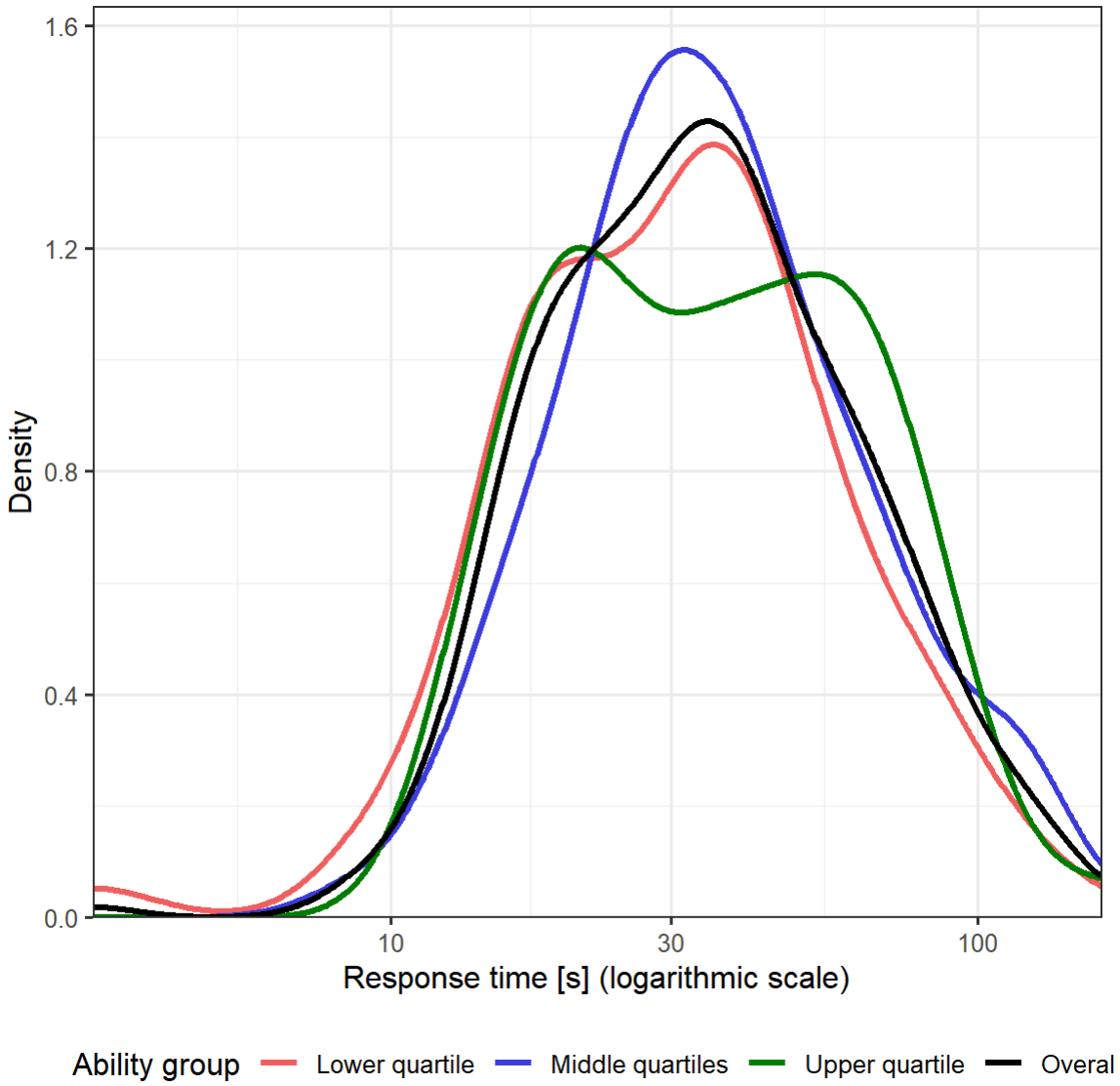
10.2.81.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.31	0.858	1.000	0.001	N/A

10.2.81.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

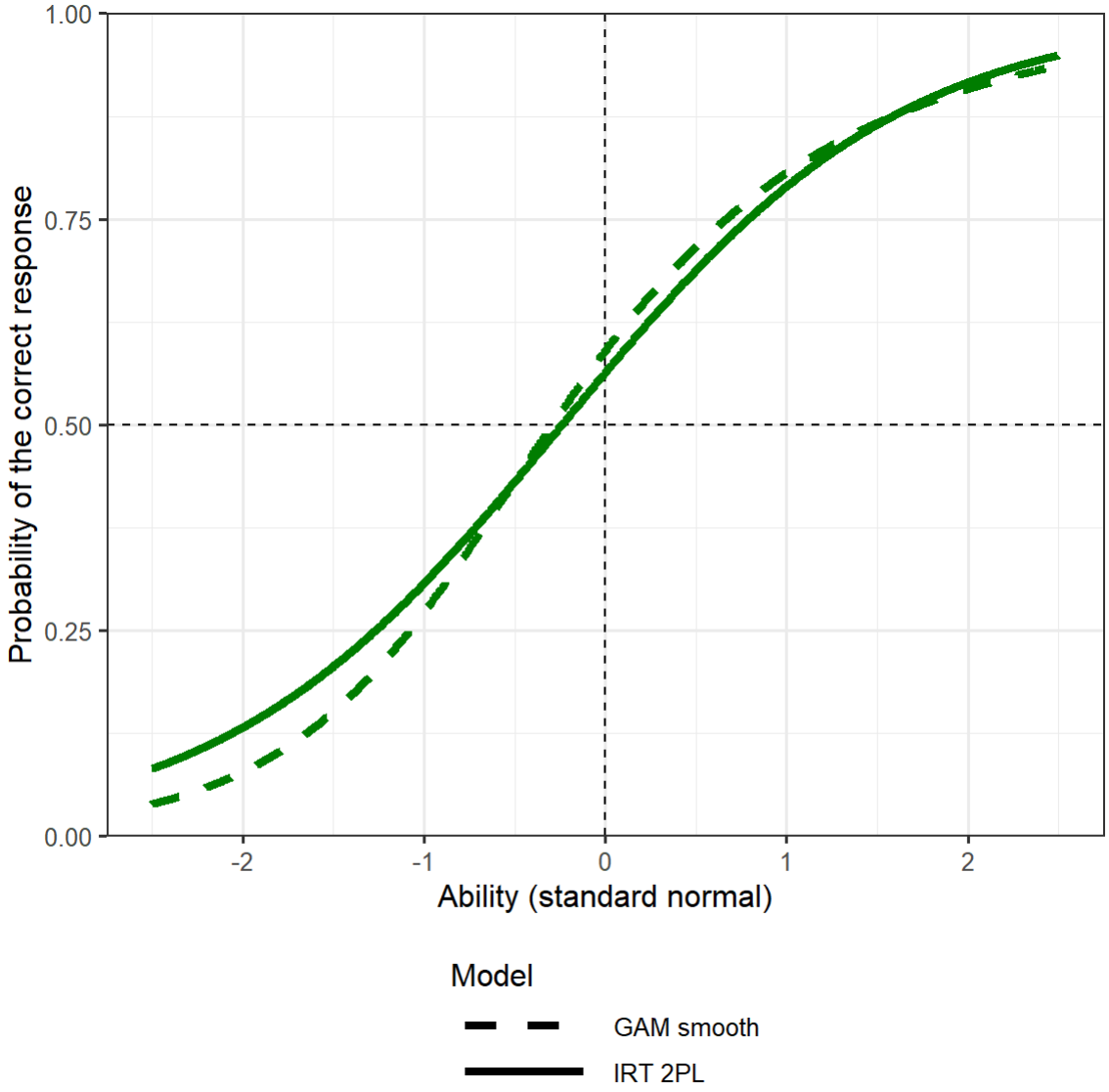
10.2.81.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	18.6	31.6	45.2	133.6	3.4	0.6
Middle quartiles	8.2	24.0	33.8	51.9	147.4	3.6	0.6
Upper quartile	12.6	20.8	35.1	58.3	162.7	3.6	0.6
Overall	3.1	21.3	33.5	51.4	162.7	3.5	0.6

10.2.82 S890009 (LLM)

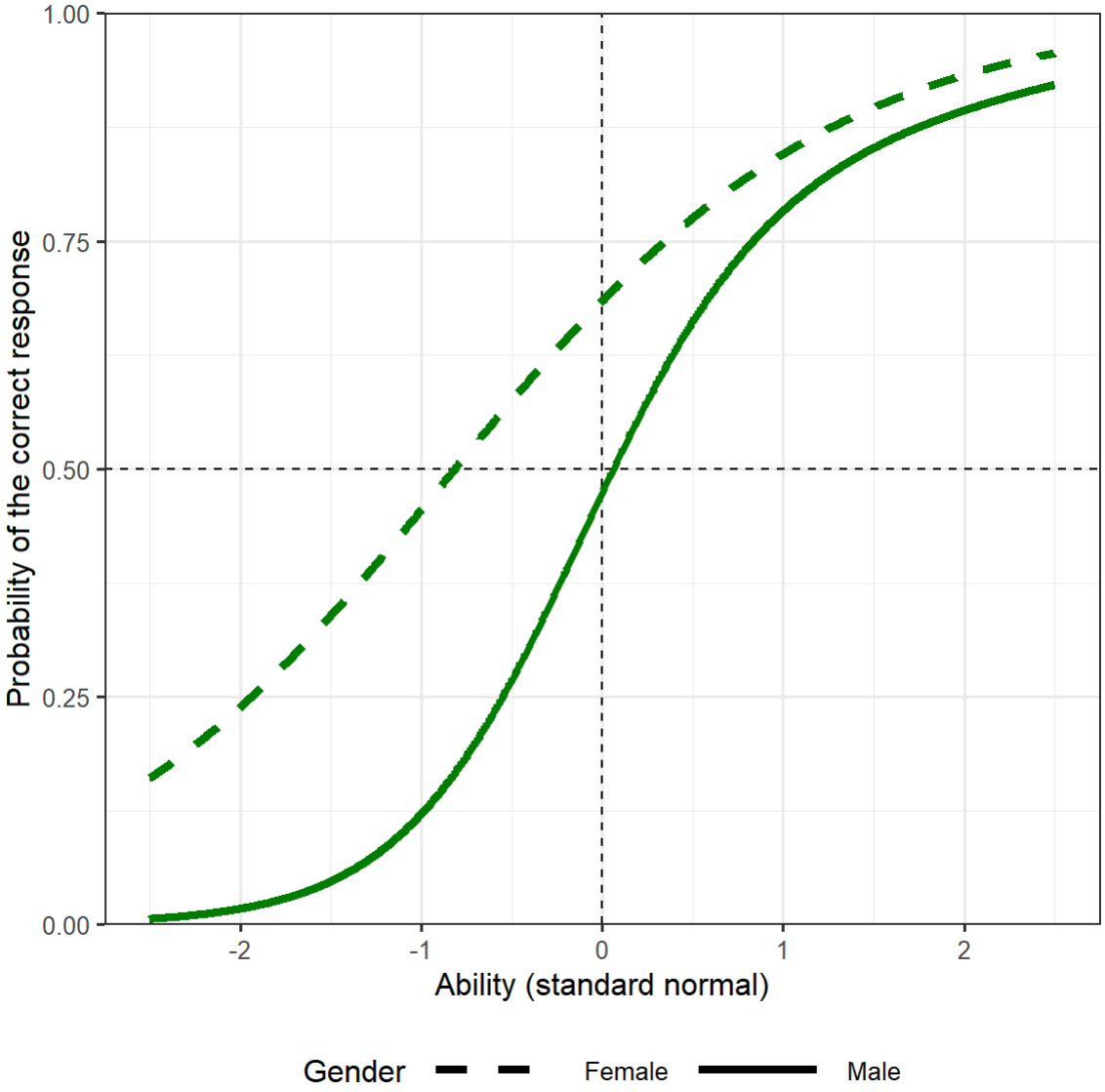
10.2.82.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	55.00%	281	LLM complex setup	CR	0.48	0.32	1.07	−0.24	0.00

10.2.82.2 Gender DIF:

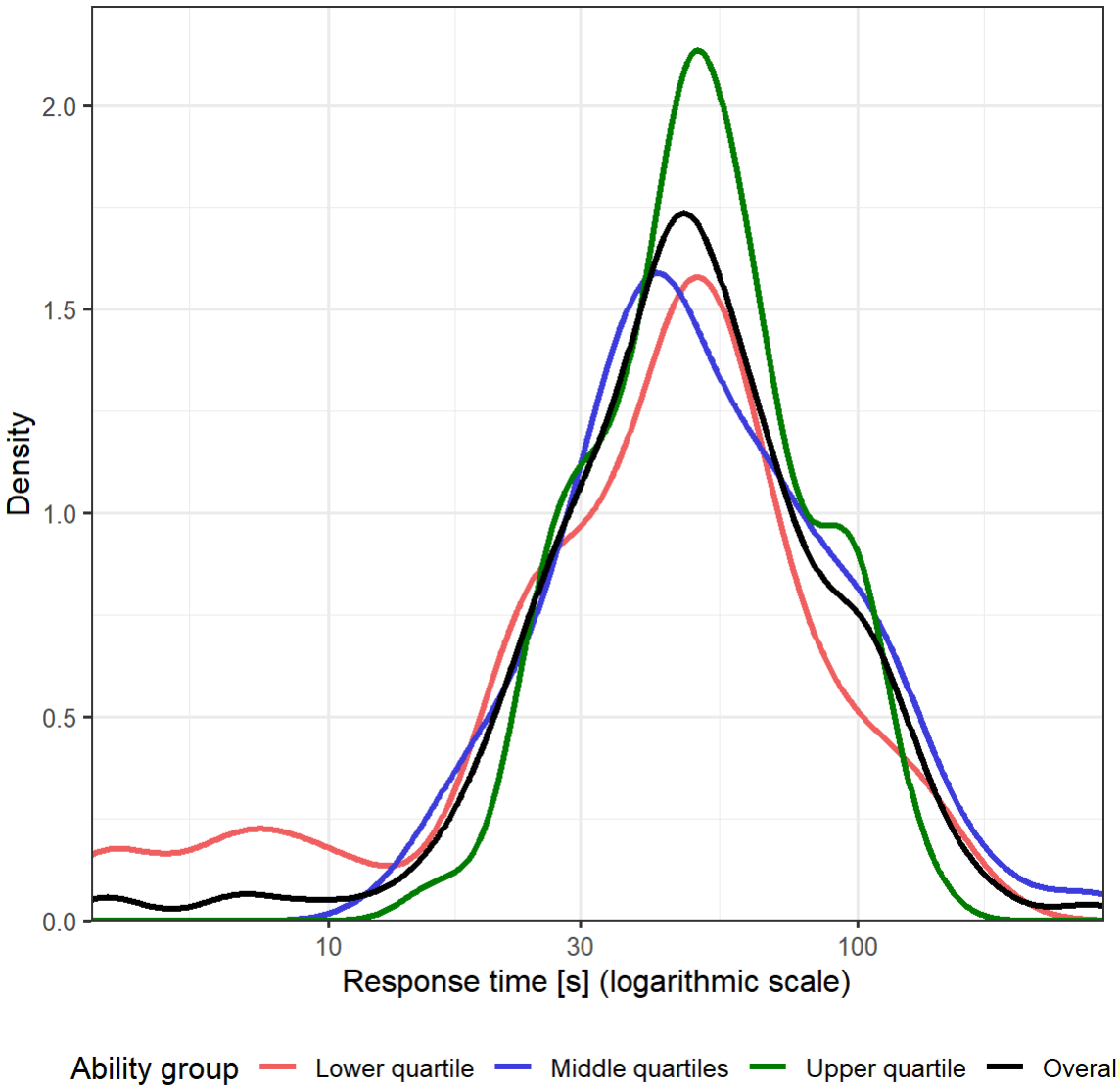
Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
19.69	0.000	0.005	0.058	moderate



10.2.82.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

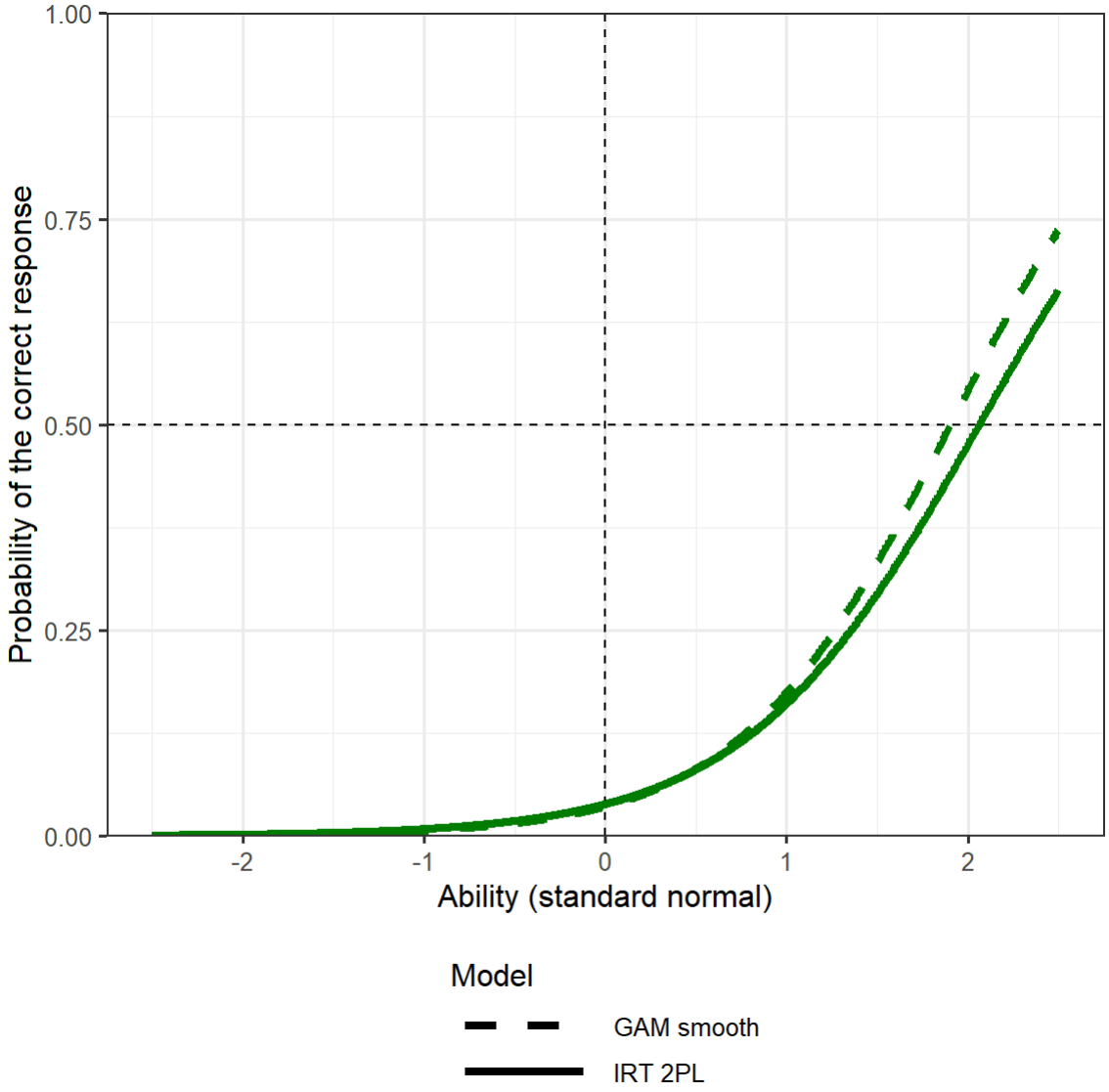
10.2.82.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.6	25.9	45.4	57.5	156.2	3.6	0.8
Middle quartiles	13.6	33.5	46.4	73.8	290.6	3.9	0.6
Upper quartile	16.2	38.8	49.6	65.6	129.2	3.9	0.4
Overall	3.6	32.6	47.7	68.6	290.6	3.8	0.6

10.2.83 S100023 (LLM)

10.2.83.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIMSS a generated items: Discrimination	Item difficulty	RMSEA
1	8.46%	319	LLM simple setup	CR	0.34	0.21	1.55	2.06	0.000

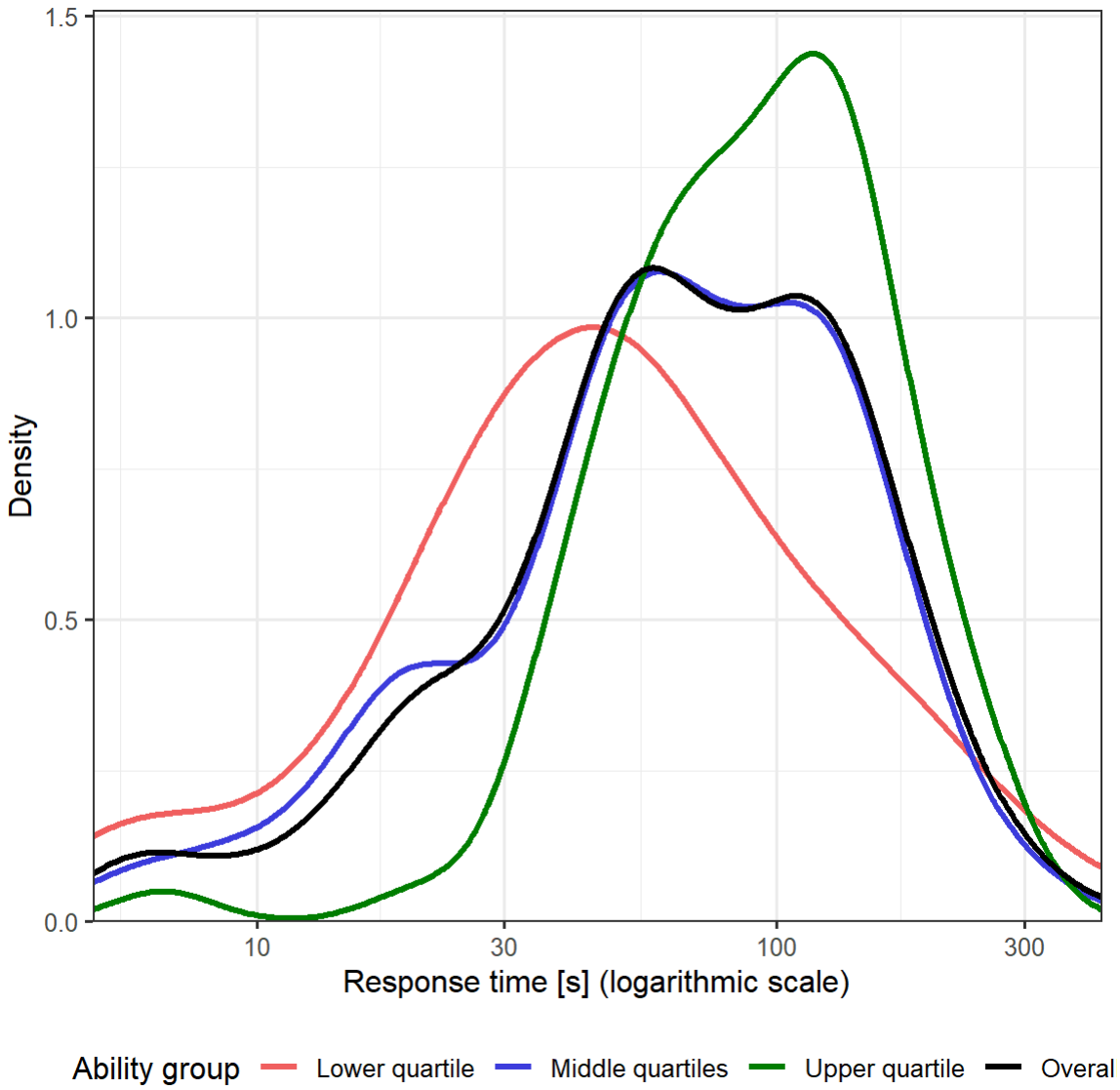
10.2.83.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.11	0.949	1.000	0.001	N/A

10.2.83.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

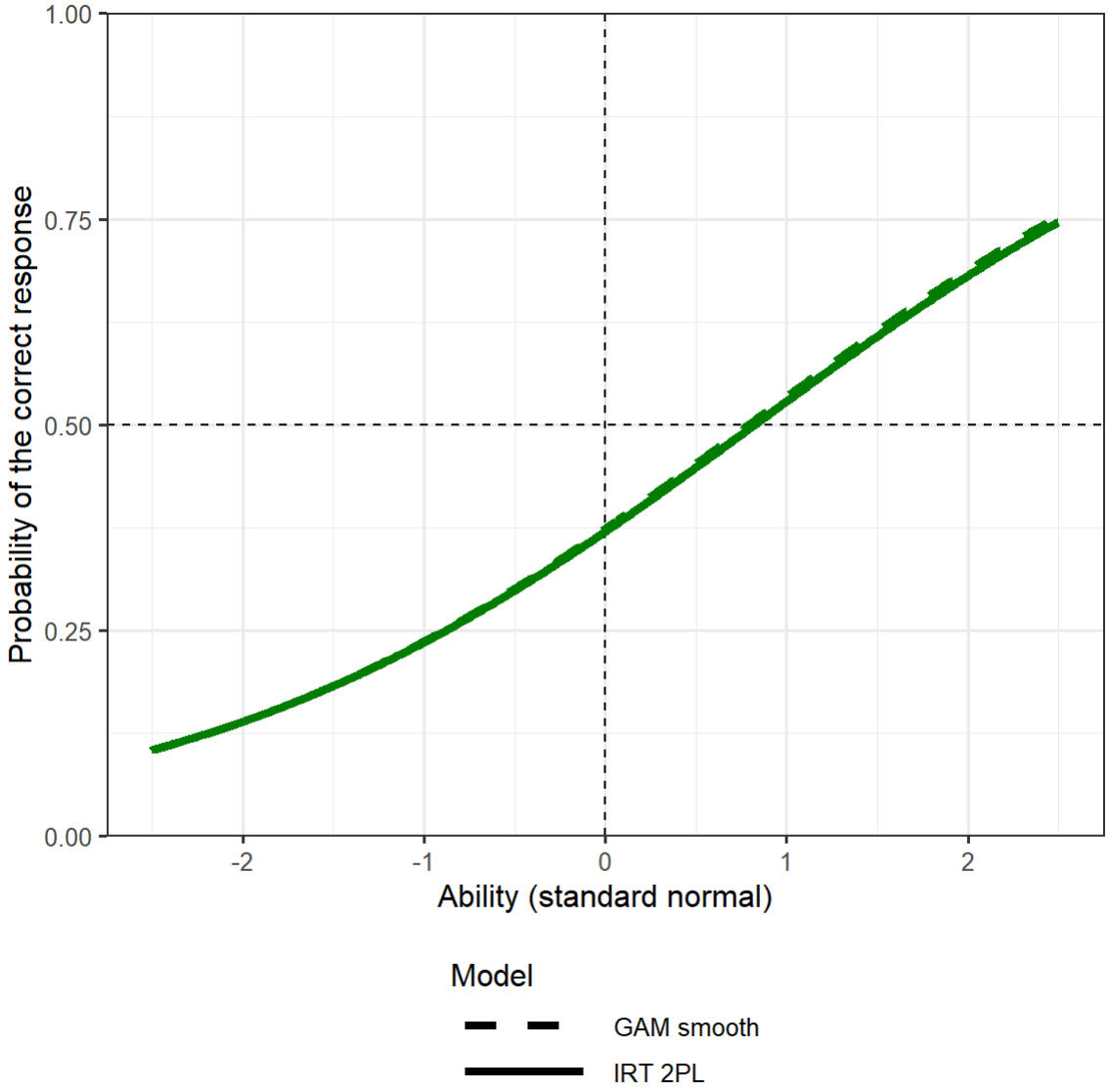
10.2.83.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	4.8	26.1	48.1	93.8	421.8	3.8	1.0
Middle quartiles	5.1	40.1	64.8	118.4	357.4	4.1	0.9
Upper quartile	6.6	58.4	92.0	135.0	285.6	4.5	0.6
Overall	4.8	40.5	66.8	120.3	421.8	4.1	0.9

10.2.84 S051201 (TIMSS)

10.2.84.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	37.38%	306	TIMSS	CR	0.29	0.28	0.65	0.82	0.036

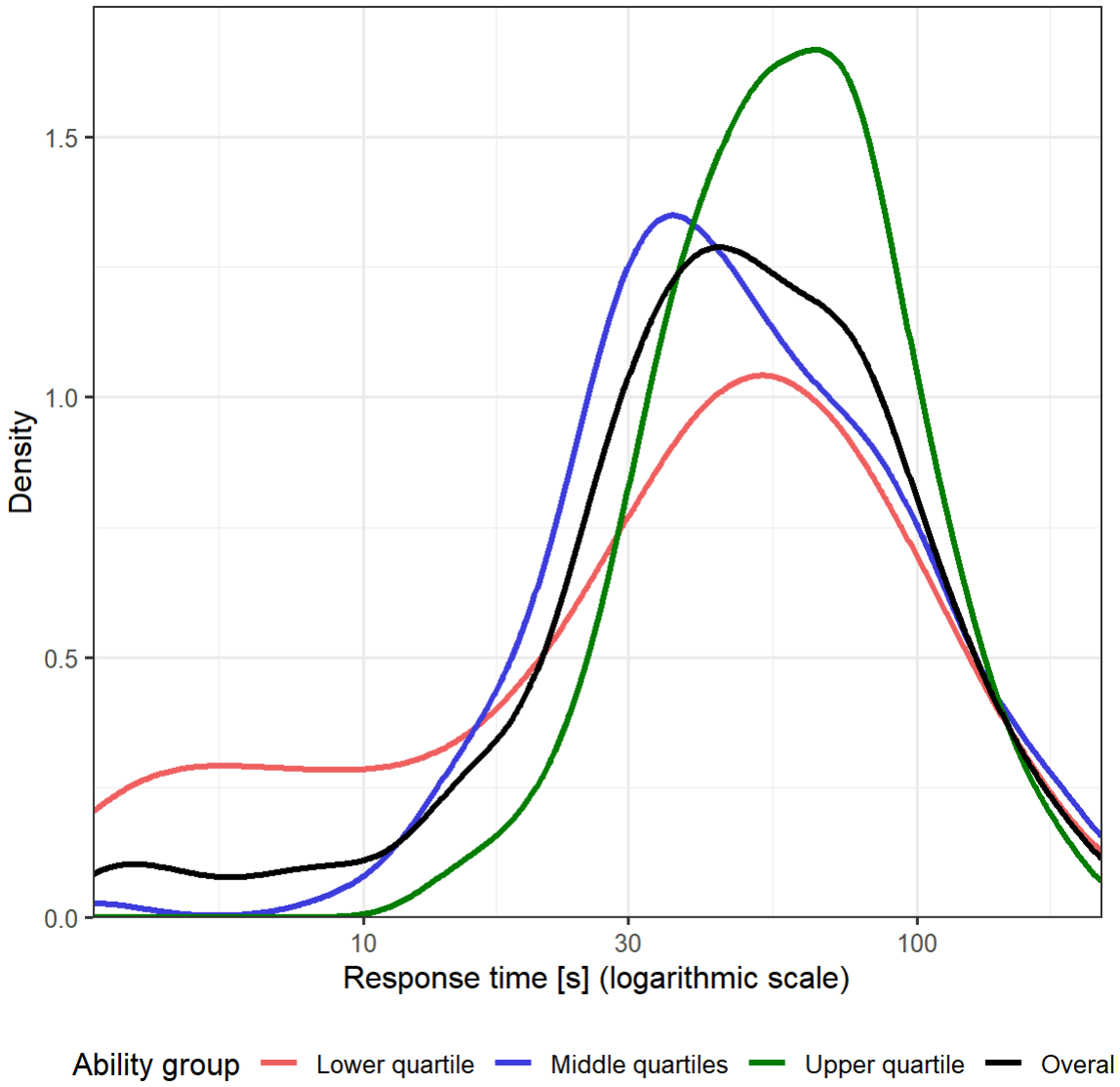
10.2.84.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.04	0.981	1.000	0.000	N/A

10.2.84.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

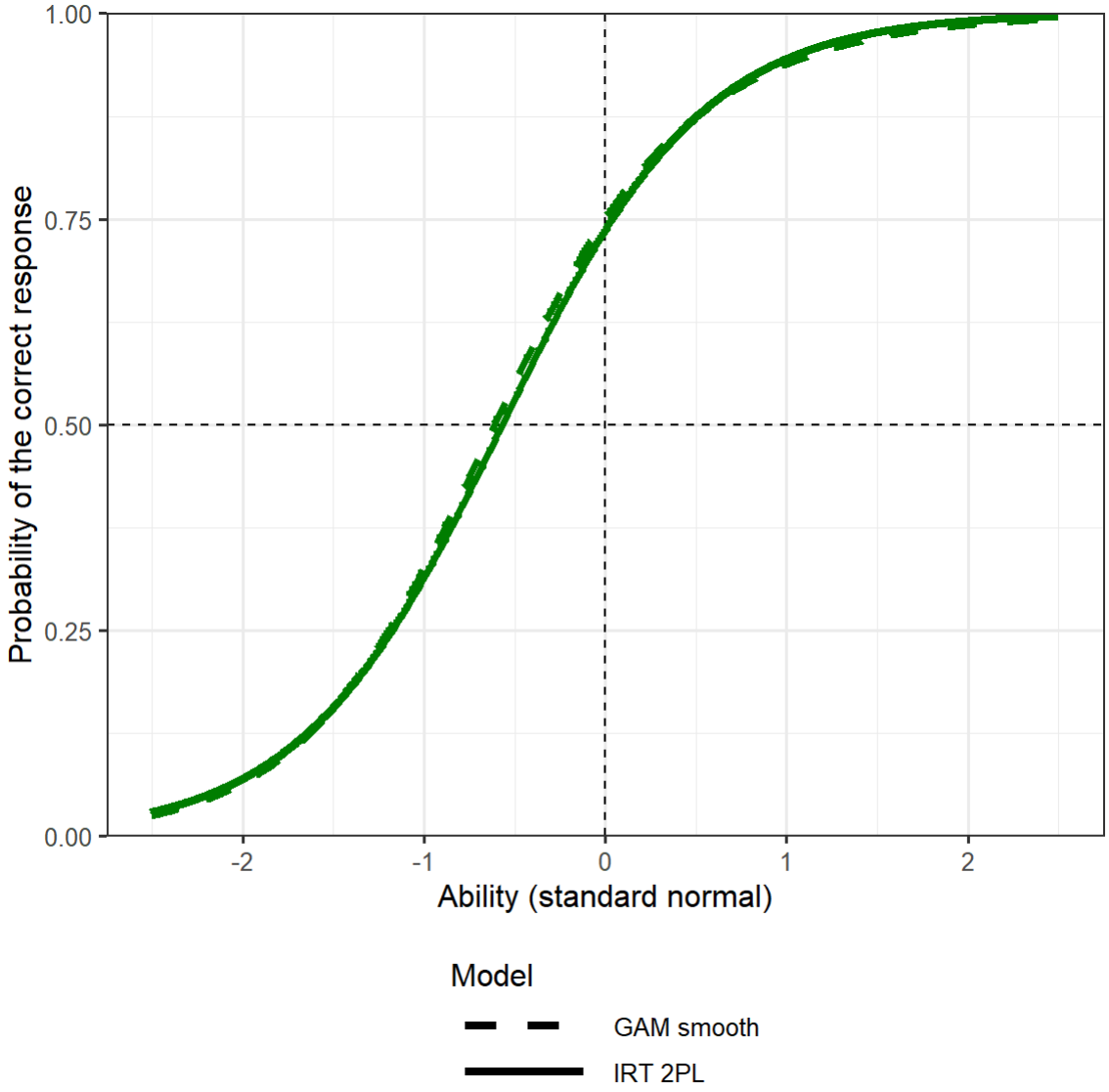
10.2.84.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.3	19.4	43.4	68.9	213.8	3.5	1.0
Middle quartiles	3.2	30.5	44.8	76.2	214.6	3.8	0.7
Upper quartile	15.1	39.7	57.3	78.0	176.5	4.0	0.5
Overall	3.2	31.0	46.6	76.1	214.6	3.8	0.8

10.2.85 SP61011 (LLM)

10.2.85.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	64.92%	306	TIMSS	CR	0.56	0.63	1.81	−0.57	0.028

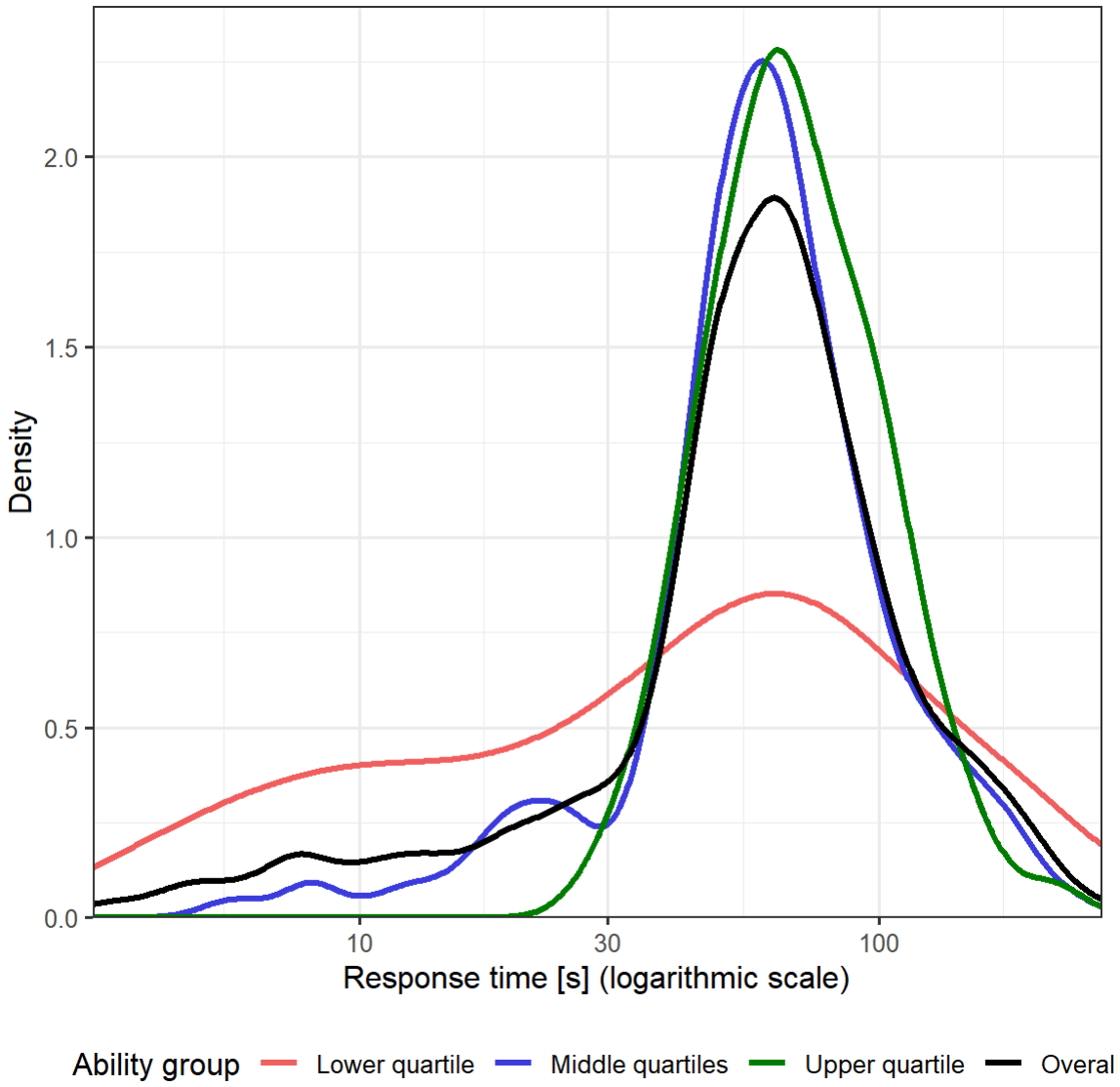
10.2.85.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
7.47	0.024	1.000	0.021	N/A

10.2.85.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

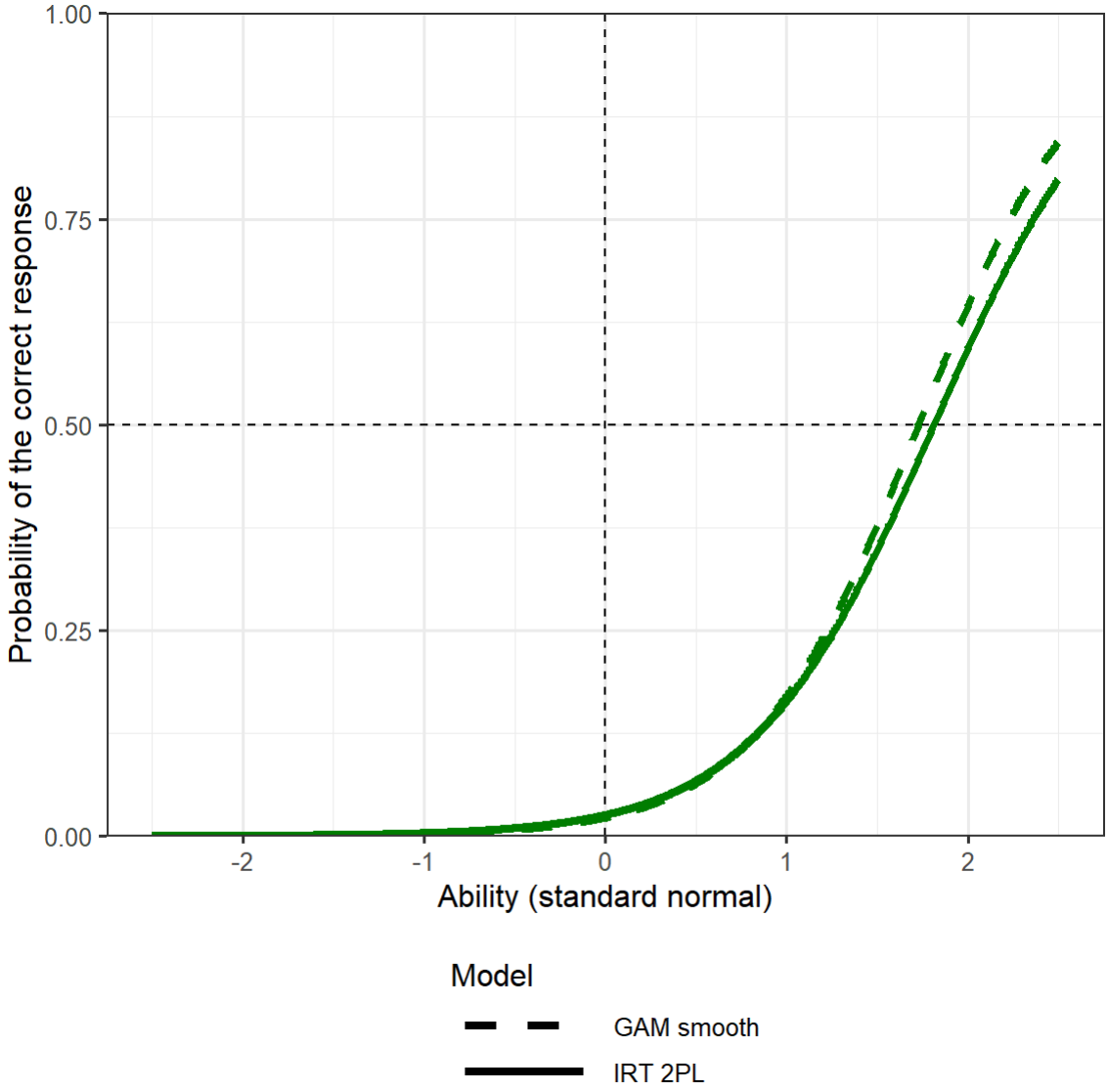
10.2.85.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	3.1	15.9	49.0	80.3	267.8	3.6	1.1
Middle quartiles	5.7	46.8	60.8	81.3	233.5	4.1	0.6
Upper quartile	29.1	52.3	66.6	89.9	214.7	4.2	0.4
Overall	3.1	43.7	61.1	83.7	267.8	4.0	0.8

10.2.86 S051164 (TIMSS)

10.2.86.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS a generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSEA
1	7.87%	306	TIMSS	CR	0.39	0.27	2.00	1.81	0.019

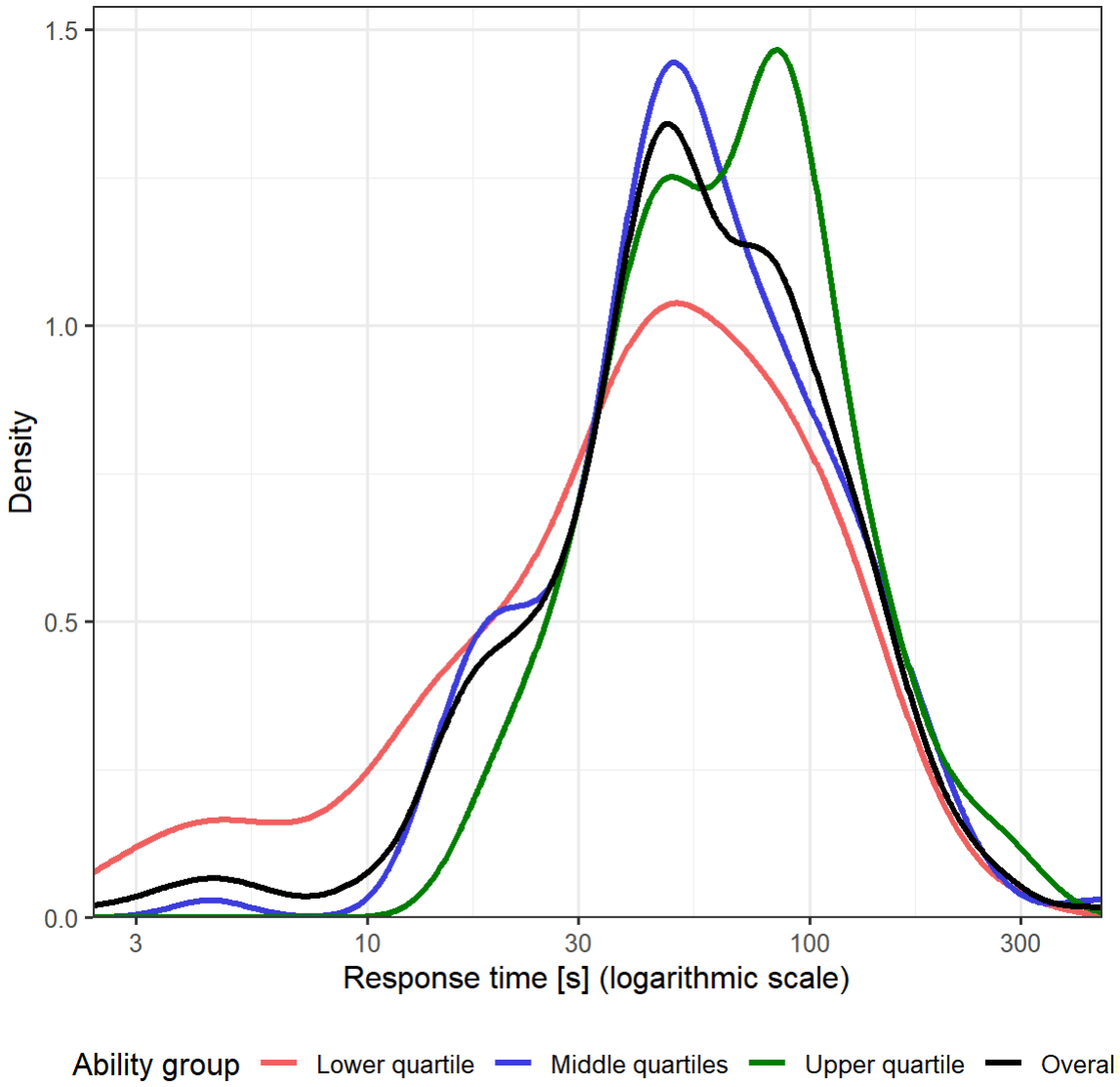
10.2.86.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.24	0.887	1.000	0.002	N/A

10.2.86.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

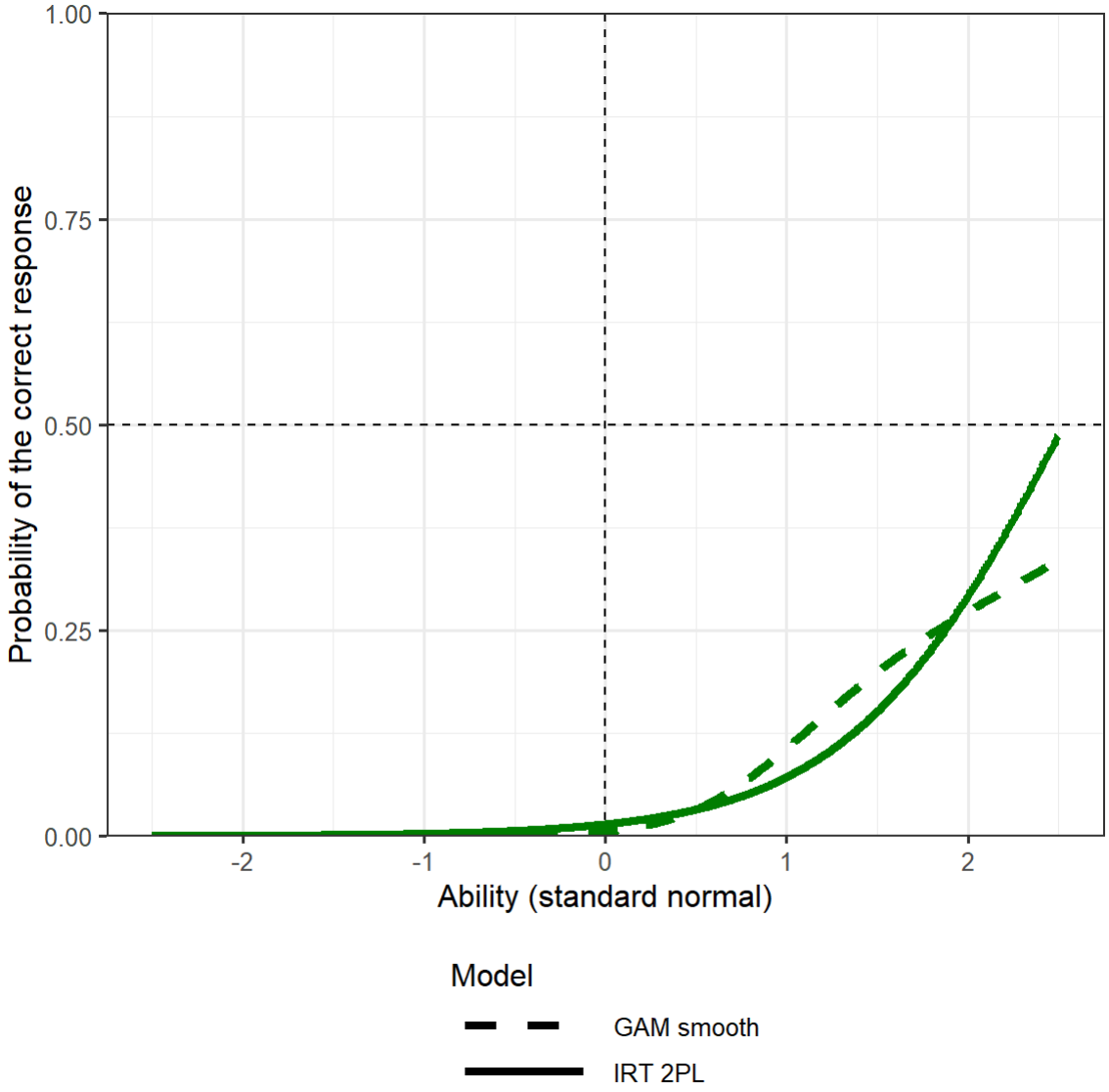
10.2.86.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	2.4	25.2	45.7	78.0	242.8	3.7	1.0
Middle quartiles	4.4	37.5	52.6	87.2	456.2	4.0	0.7
Upper quartile	18.8	43.8	69.2	93.0	297.0	4.2	0.6
Overall	2.4	35.6	52.6	87.4	456.2	3.9	0.8

10.2.87 S890005 (TIMSS)

10.2.87.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMSS generated items		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	RMSE
1	4.08%	319	LLM complex setup	CR	0.27	0.23	1.67	2.53	0.01

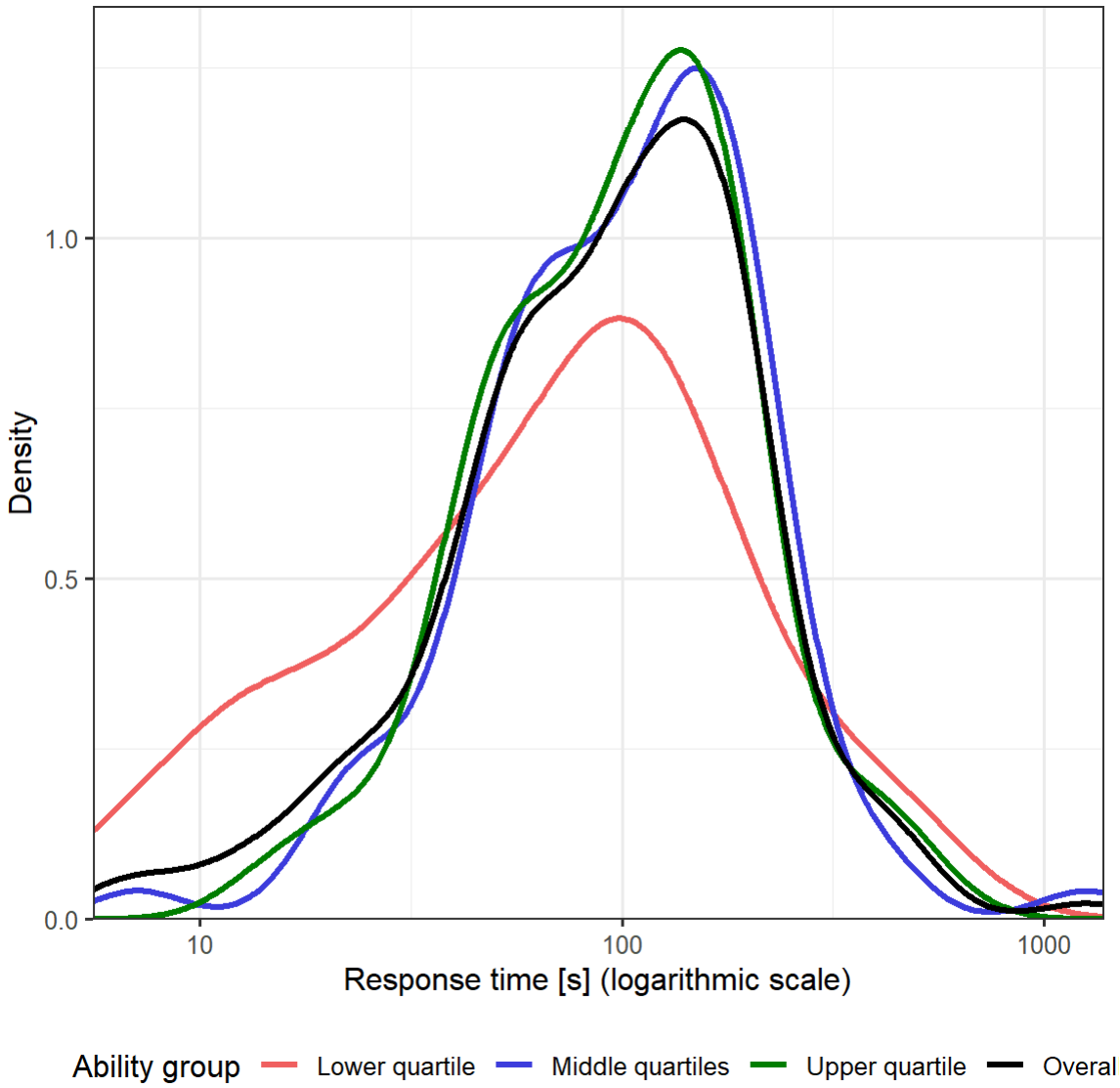
10.2.87.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
0.42	0.810	1.000	0.004	N/A

10.2.87.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

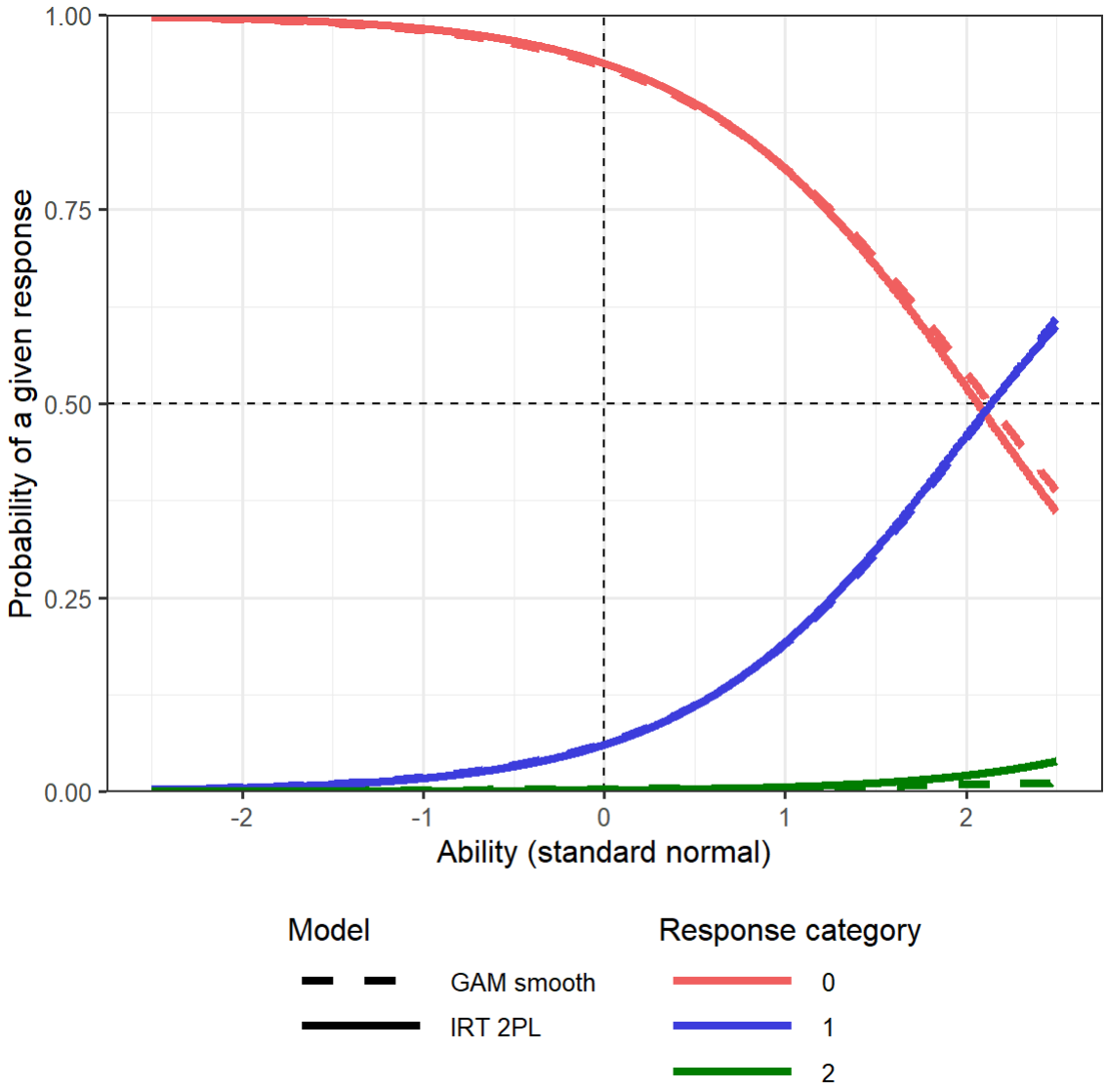
10.2.87.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	5.6	33.7	84.0	131.1	541.0	4.2	1.1
Middle quartiles	6.6	61.9	109.3	172.6	1,381.2	4.6	0.8
Upper quartile	13.8	58.9	104.2	159.6	541.8	4.6	0.7
Overall	5.6	54.2	100.3	161.5	1,381.2	4.5	0.9

10.2.88 S890006 (TIMSS)

10.2.88.1 Item fit:



Max. points	Easiness	Number of obs.	Source	Format	$r_{score,\theta}$ All items	$r_{score,\theta}$ TIMSS items only	IRT model including both TIM Discrimination	IRT model including both TIM Item difficulty	IRT model including both TIM Categ difficu
2	5.54%	281	LLM complex setup	CR	0.31	0.24	1.32	3.50	2

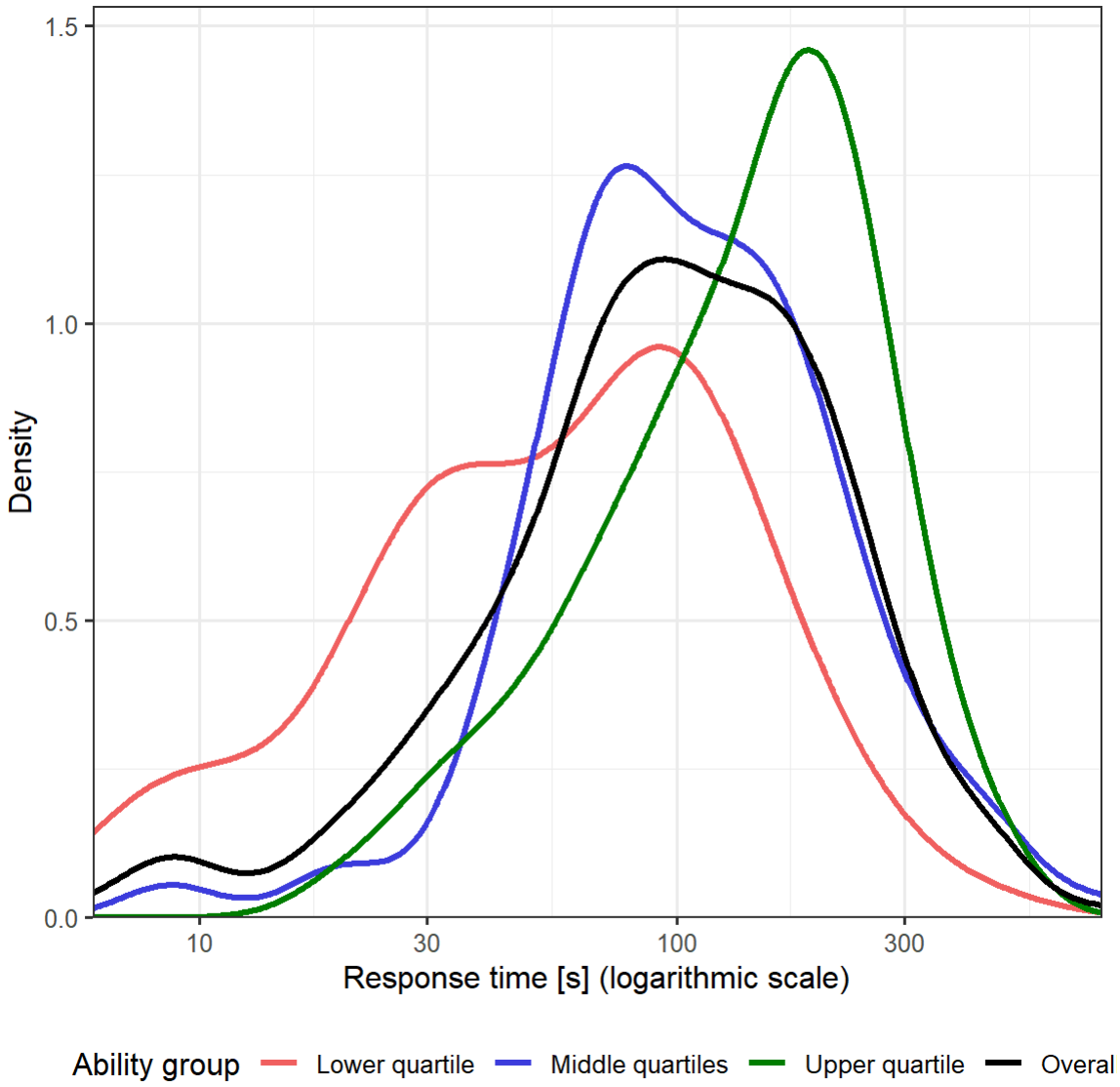
10.2.88.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
1.75	0.417	1.000	0.010	N/A

10.2.88.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

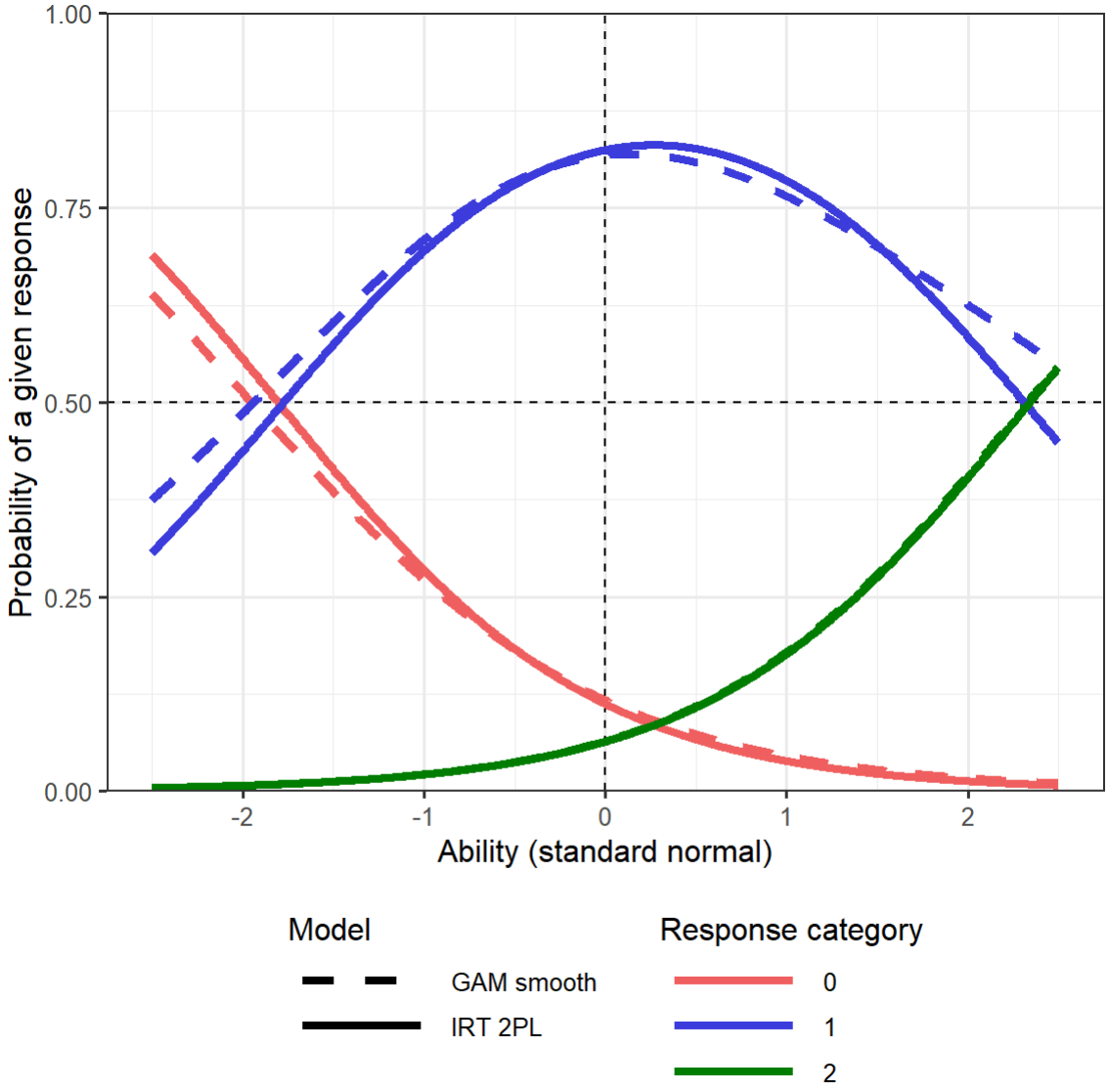
10.2.88.4 Answering time:



	Time [s]					log(time)	
Ability group	Min	Q1	Median	Q3	Max	Mean	SD
Lower quartile	6.0	28.8	65.4	103.4	419.1	4.0	0.9
Middle quartiles	8.2	66.2	105.8	168.4	775.0	4.6	0.7
Upper quartile	20.1	92.5	159.0	220.0	456.0	4.9	0.7
Overall	6.0	60.5	102.7	169.7	775.0	4.5	0.8

10.2.89 S041183 (TIMSS)

10.2.89.1 Item fit:



					$r_{score,\theta}$		IRT model including both TIMS		
Max. points	Easiness	Number of obs.	Source	Format	All items	TIMSS items only	Discrimination	Item difficulty	Catego difficul
2	46.39%	306	TIMSS	CR	0.41	0.42	1.15	0.27	−1.8

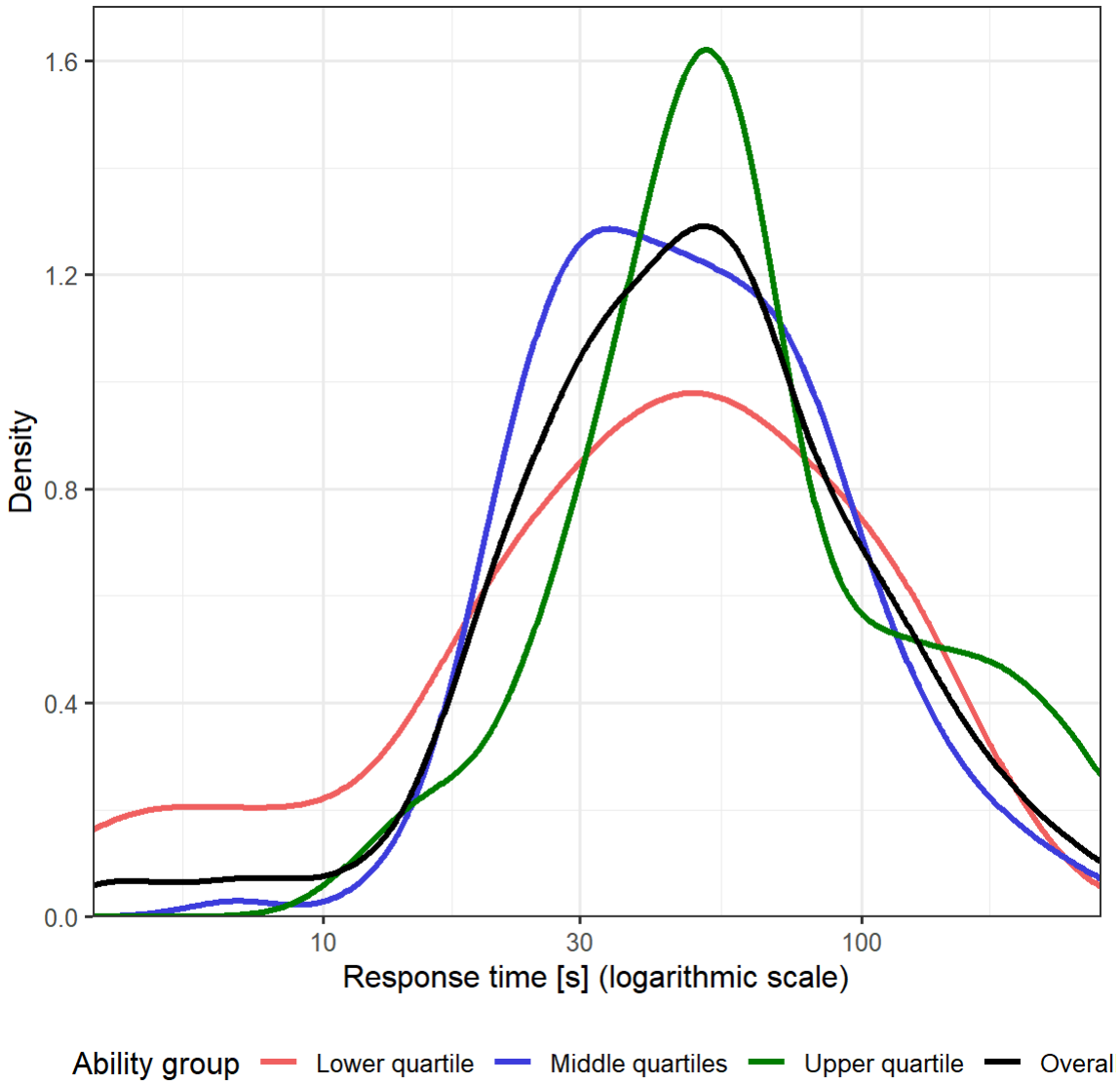
10.2.89.2 Gender DIF:

Statistic	P-value	Holm's corrected p-value	ΔNagelkerke's R^2	Effect size (Jodoin & Gierl)
7.93	0.019	1.000	0.032	N/A

10.2.89.3 Frequency of answers:

The item is open-ended. The distribution of answers can not be shown.

10.2.89.4 Answering time:


1.6
1.2
0.8
0.4
0.0
Density
10
30
100
Response time [s] (logarithmic scale)
Ability group
Lower quartile
Middle quartiles
Upper quartile
Overal