Test and Item Analysis Report

HCI dataset

Joe Doe

Report created on **December 13, 2018**

Contents

Introduction

This report was created by ShinyItemAnalysis application version 1.2.9. ShinyItemAnalysis provides test and item analysis and it is available on CRAN and also online.

To cite ShinyItemAnalysis application in publications, please, use:

[1] Martinková P., & Drabinová A. (2018) ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*. Accepted.

[2] Martinková P., Drabinová A., Leder O., & Houdek J. (2018) ShinyItemAnalysis: Test and item analysis via shiny. R package version 1.2.9. https://CRAN.R-project.org/package=ShinyItemAnalysis

[3] Martinková P., Drabinová A., & Houdek J. (2017) ShinyItemAnalysis: Analýza přijímacích a jiných znalostních či psychologických testů [ShinyItemAnalysis: Analyzing admission and other educational and psychological tests, in Czech]. *TESTFÓRUM, 6*(9), 16-35. doi:10.5817/TF2017-9-129

This project was supported by Czech Science Foundation under grant number GJ15-15856Y.

Summary

Total scores

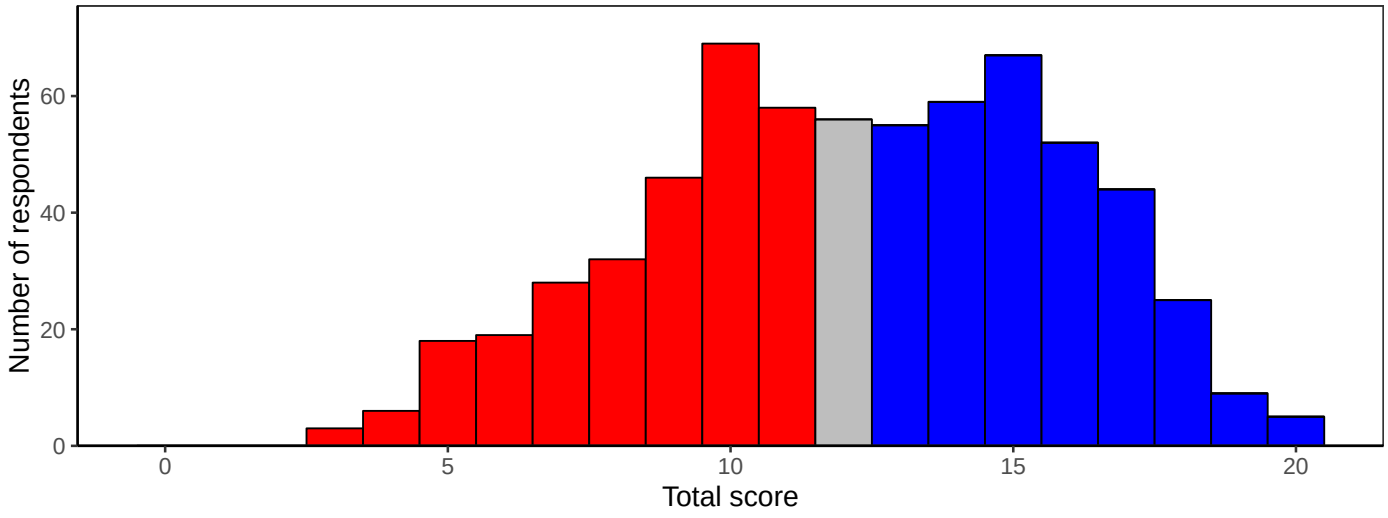
Summary table of total scores

The table below summarizes basic characteristics of total scores including minimum and maximum, mean, median, standard deviation, skewness and kurtosis. The kurtosis here is estimated by a sample kurtosis $\frac{m_4}{m_2^2}$, where m_4 is the fourth central moment and m_2 is the second central moment. The skewness is estimated by a sample skewness $\frac{m_3}{m_2^{3/2}}$, where m_3 is the third central moment. The kurtosis for normally distributed scores is near the value of 3 and the skewness is near the value of 0.

Min	Max	Mean	Median	SD	Skewness	Kurtosis
3	20	12.21	12.00	3.64	-0.20	2.35

Histogram of total scores

For a selected cut-score, the blue part of a histogram shows respondents with a total score above the cut-score, the grey column shows respondents with a total score equal to the cut-score and the red part of a histogram shows respondents below the cut-score.



Scoring

Summary table of standard scores

The total score, also known as the raw score, is the total number of correct answers. It can be used to compare an individual score to a norm group, e.g. if the mean is 12, then an individual score can be compared to see if it is below or above this average. The percentile indicates the value below which a percentage of observations falls, e.g. an individual score at the 80th percentile means that the individual score is the same or higher than the scores of 80% of all respondents. The success rate is the percentage of correct answers, e.g. if the maximum points of a test is equal to 20 and an individual score is 12 then the success rate is 12/20 = 0.6, i.e. 60%. The Z-score, or the standardized score, is a linear transformation of the total score with a mean of 0 and with a variance of 1. If X is the total score, M is its mean and SD is its standard deviation then Z-score = (X - M) / SD. The T-score is the transformed Z-score with a mean of 50 and a standard deviation of 10. If Z is Z-score then T-score = (Z * 10) + 50.

Total score	Percentile	Success rate	Z-score	T-score
3	0.00	15	-2.53	24.69
4	0.01	20	-2.26	27.44
5	0.04	25	-1.98	30.19
6	0.07	30	-1.71	32.93
7	0.11	35	-1.43	35.68
8	0.16	40	-1.16	38.43
9	0.23	45	-0.88	41.18
10	0.34	50	-0.61	43.92
11	0.43	55	-0.33	46.67
12	0.51	60	-0.06	49.42
13	0.60	65	0.22	52.17
14	0.69	70	0.49	54.91
15	0.79	75	0.77	57.66
16	0.87	80	1.04	60.41
17	0.94	85	1.32	63.16
18	0.98	90	1.59	65.90
19	0.99	95	1.87	68.65
20	1.00	100	2.14	71.40

More complex estimates of ability are provided in the IRT section.

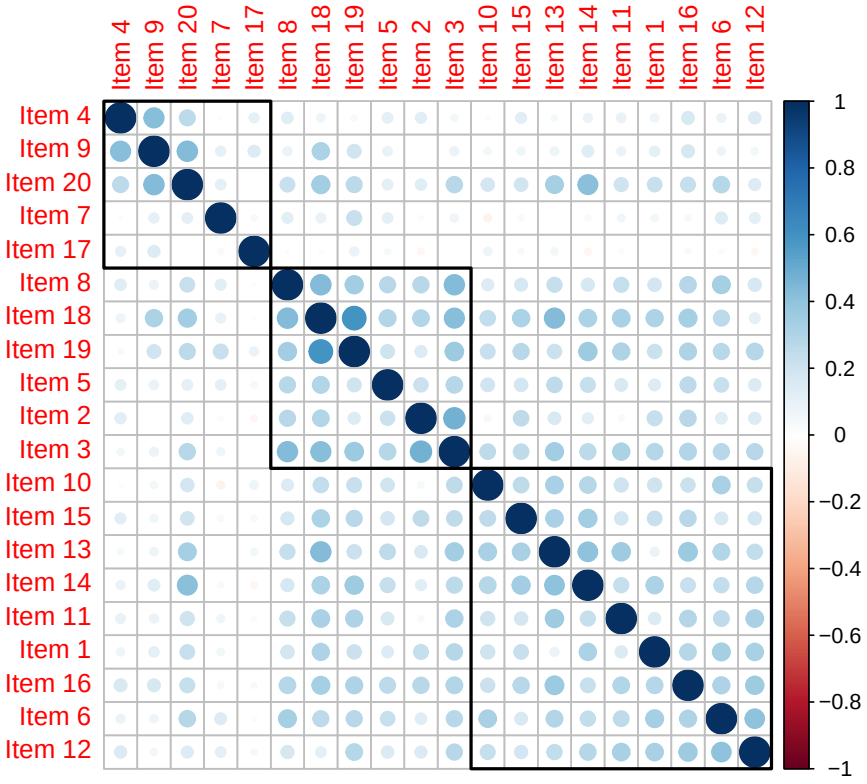
Validity

Correlation structure

Correlation heat map

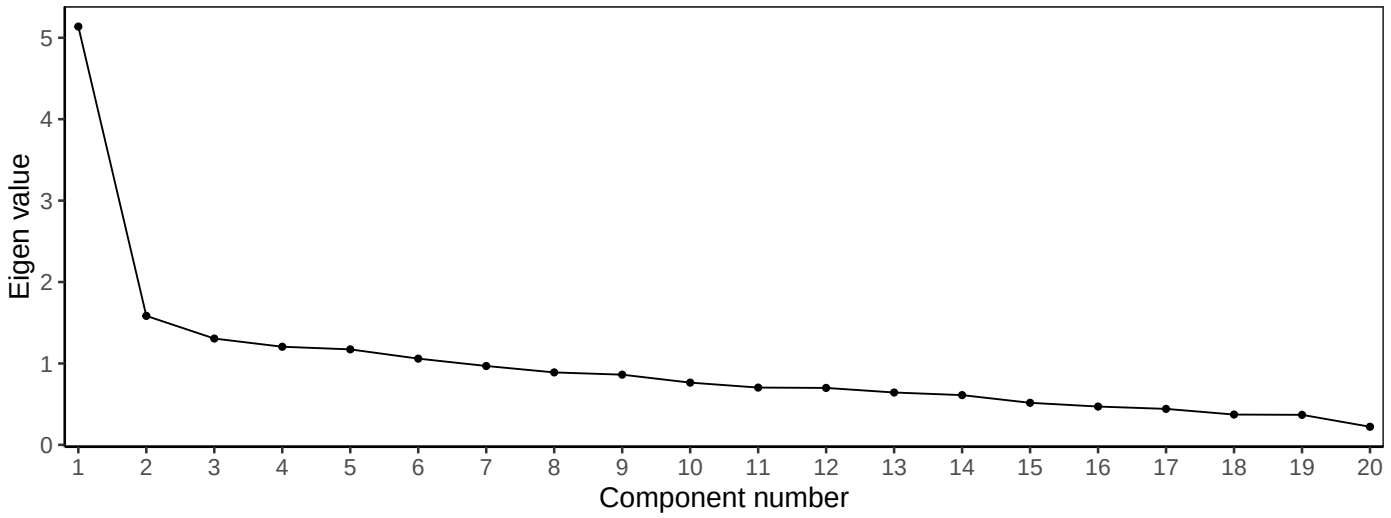
A correlation heat map displays polychoric correlations of items. Polychoric/tetrachoric correlation between two ordinal/binary variables is calculated from their contingency table, under the assumption that the ordinal variables dissect continuous latent variables that are bivariate normal.

A correlation heat map is reordered using hierarchical clustering with Ward's linkage method. The number of highlighted clusters is 3 .



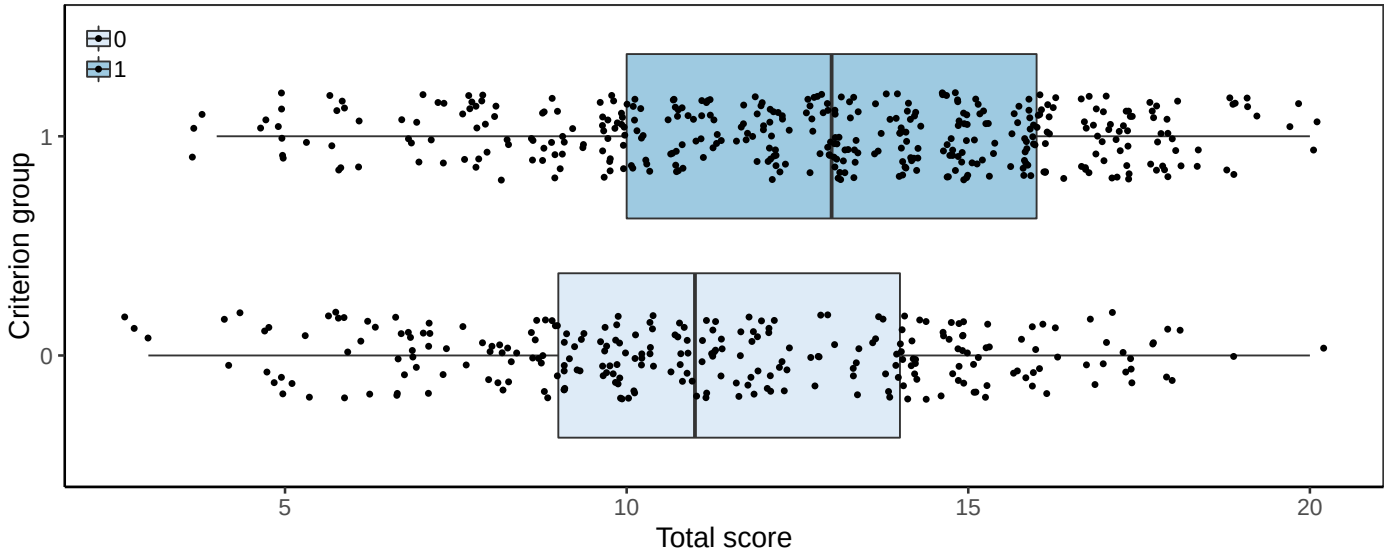
Scree plot

A scree plot displays the eigenvalues associated with a component or a factor in descending order.



Predictive validity

A test for association between the total score and the criterion variable is based on Spearman's rho. The null hypothesis is that the correlation is 0. Results: rho = 0.22 (p-value <0.01). Interpretation: The p-value is less than 0.05, thus we reject the null hypothesis; total score and the criterion variable are positively correlated.

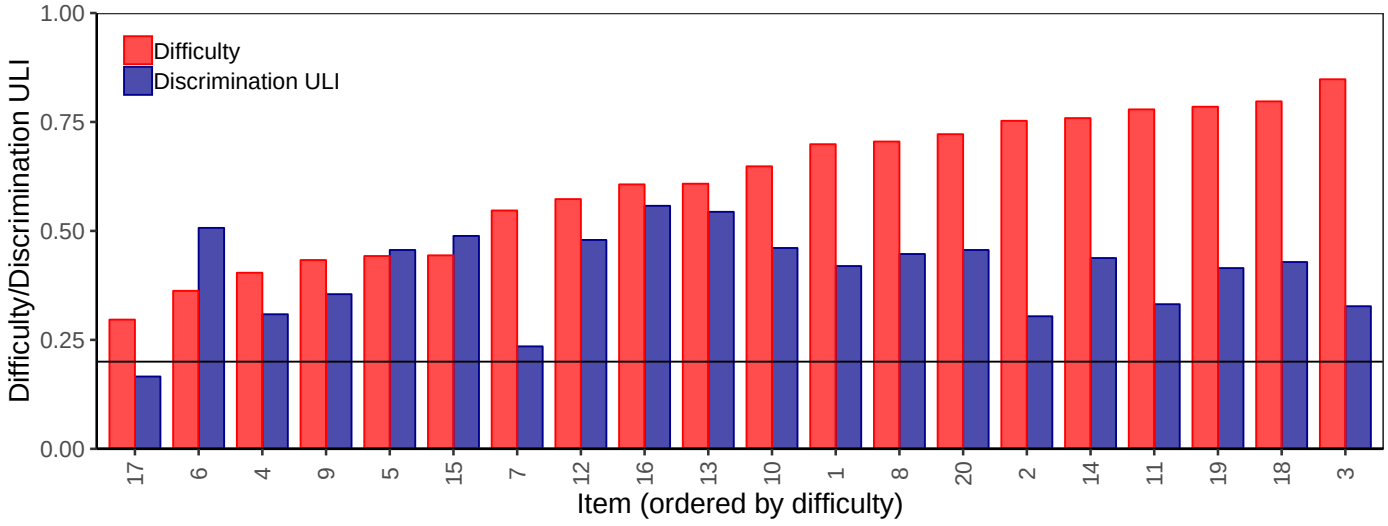


Traditional item analysis

Item analysis

Difficulty/Discrimination plot

Difficulty (red) of items is estimated as a percentage of the respondents who answered correctly to that item. Discrimination (blue) is described by the difference of the percent correct in the upper and lower third of respondents (Upper-Lower Index, ULI). By a rule of thumb, it should not be lower than 0.2 (borderline in the plot), except for very easy or very difficult items.



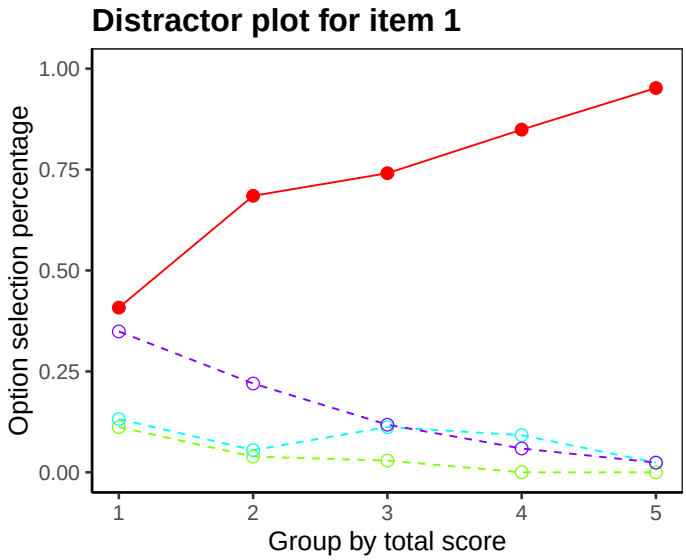
Traditional item analysis table

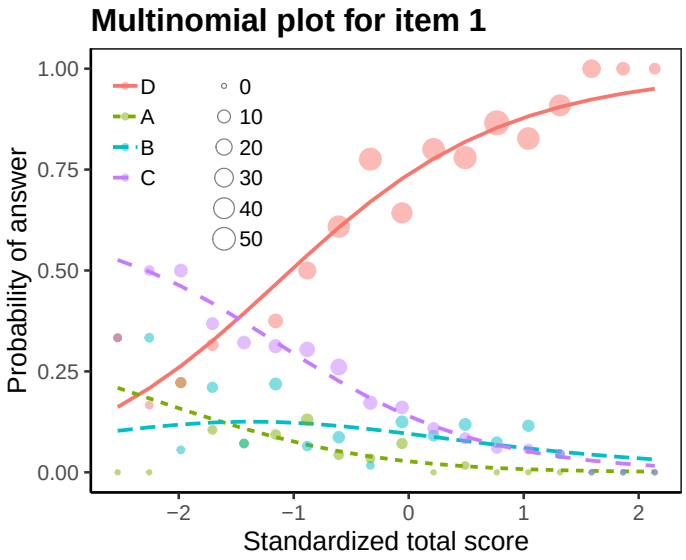
Diff. - Difficulty of an item is estimated as the percent of respondents who correctly answered that item. SD - standard deviation, ULI - Upper-Lower Index calculated as the difference in the percent correct between the upper and lower third of respondents, RIT - Pearson correlation between an item and the total score, RIR - Pearson correlation between an item and the rest of the items, Alpha Drop - Cronbach's alpha of a test without a given item, gULI - generalized ULI calculated as the difference between the difficulty recorded in the groups selected by the user.

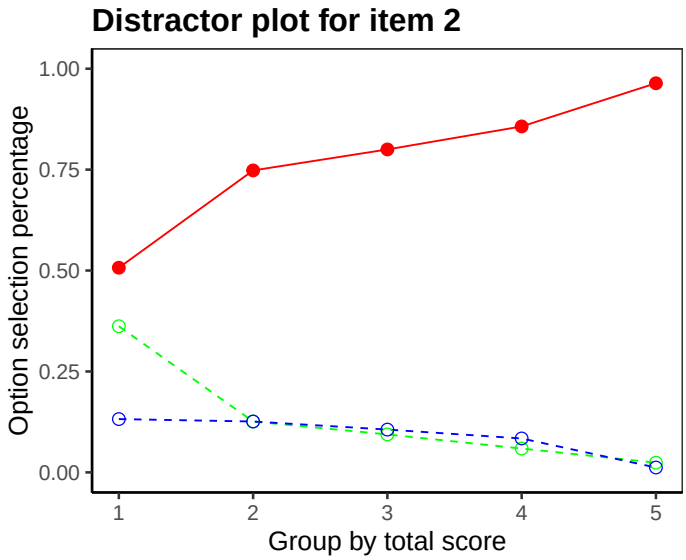
Item	Diff.	SD	ULI	RIT	RIR	Alpha Drop	gULI
1	0.70	0.46	0.42	0.40	0.29	0.70	0.42
2	0.75	0.43	0.30	0.33	0.22	0.71	0.30
3	0.85	0.36	0.33	0.44	0.35	0.70	0.33
4	0.40	0.49	0.31	0.30	0.17	0.72	0.31
5	0.44	0.50	0.46	0.40	0.28	0.71	0.46
6	0.36	0.48	0.51	0.46	0.34	0.70	0.51
7	0.55	0.50	0.24	0.24	0.10	0.72	0.24
8	0.71	0.46	0.45	0.44	0.33	0.70	0.45
9	0.43	0.50	0.35	0.34	0.21	0.71	0.35
10	0.65	0.48	0.46	0.38	0.26	0.71	0.46
11	0.78	0.42	0.33	0.39	0.29	0.70	0.33
12	0.57	0.50	0.48	0.44	0.32	0.70	0.48
13	0.61	0.49	0.54	0.48	0.36	0.70	0.54
14	0.76	0.43	0.44	0.44	0.34	0.70	0.44
15	0.44	0.50	0.49	0.42	0.30	0.70	0.49
16	0.61	0.49	0.56	0.49	0.38	0.70	0.56
17	0.30	0.46	0.17	0.17	0.04	0.73	0.17
18	0.80	0.40	0.43	0.51	0.42	0.69	0.43
19	0.78	0.41	0.41	0.46	0.37	0.70	0.41
20	0.72	0.45	0.46	0.45	0.34	0.70	0.46

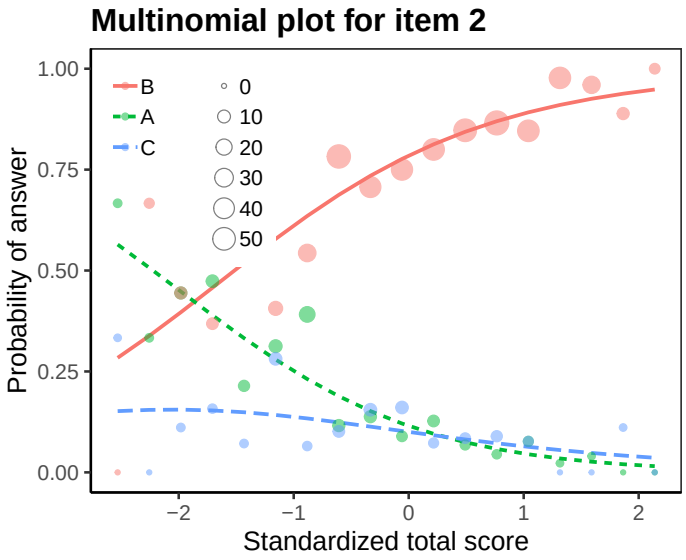
Distractor analysis

Respondents are divided into a selected number of groups by their total score. Subsequently, the percentage of respondents in each group who selected a given answer (correct answer or distractor) is displayed. The correct answer should be selected more often by the respondents with a higher total score than by those with a lower total score, i.e. the solid line should be increasing. The distractor should work in the opposite direction, i.e. the dotted lines should be decreasing.

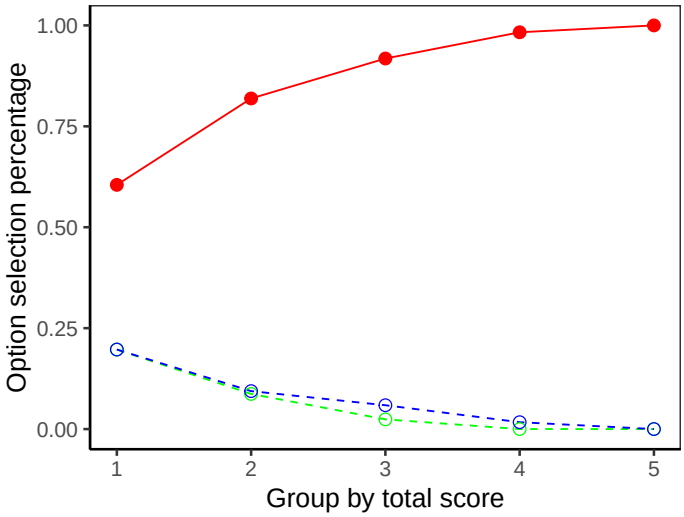




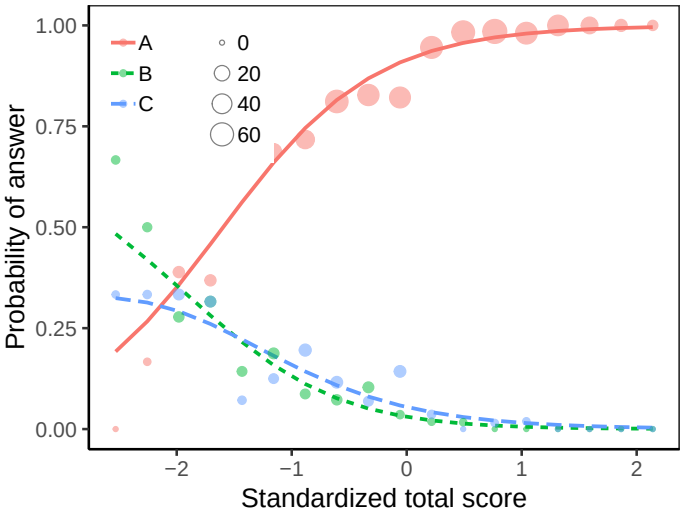




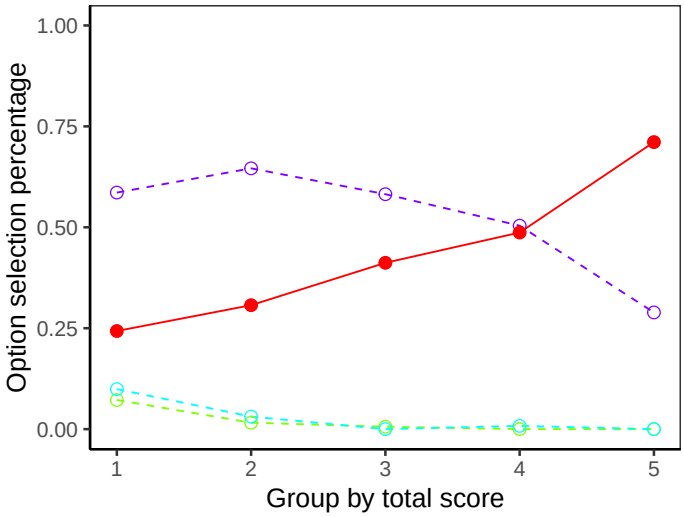
Distractor plot for item 3



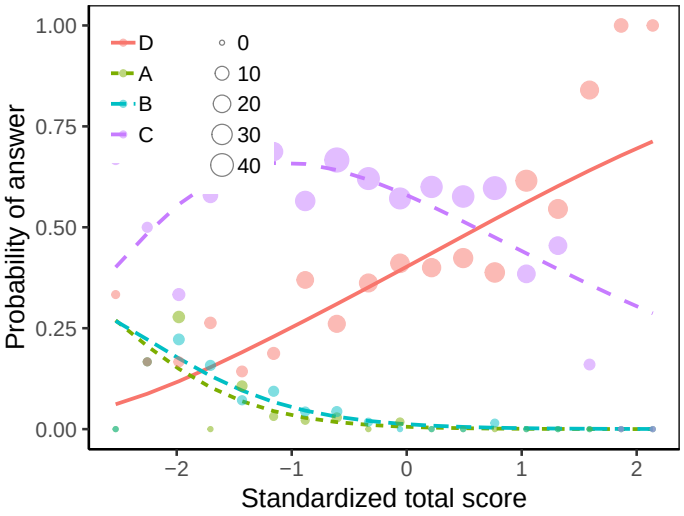
Multinomial plot for item 3



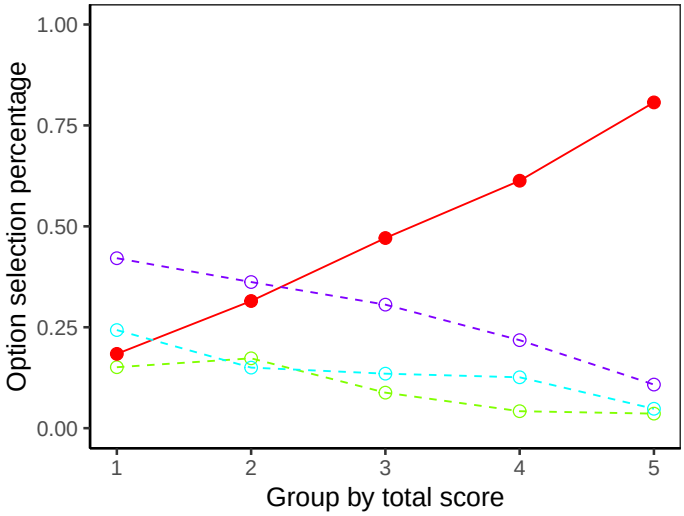
Distractor plot for item 4



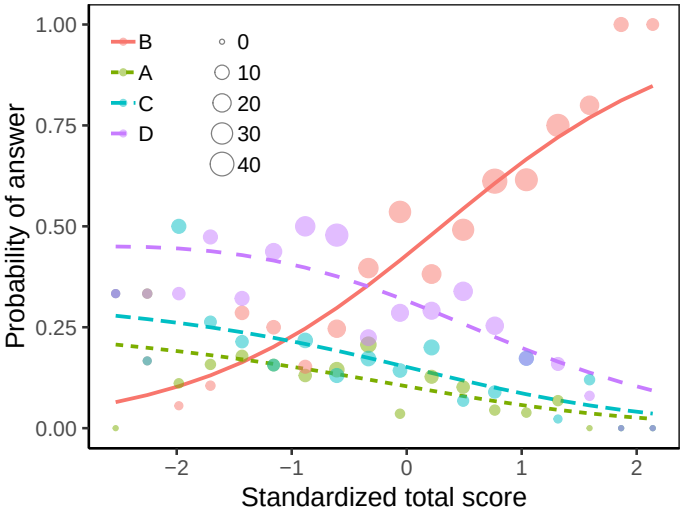
Multinomial plot for item 4



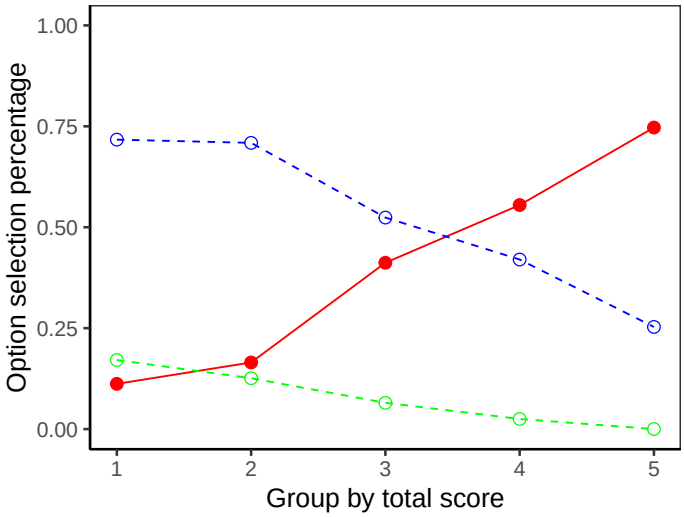
Distractor plot for item 5



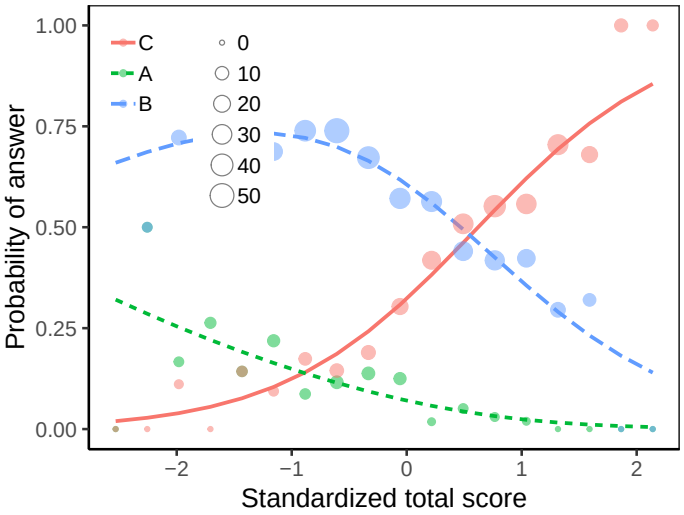
Multinomial plot for item 5



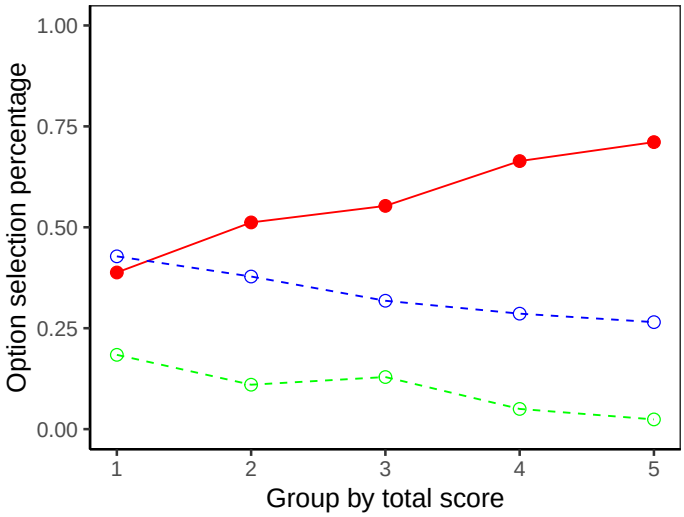
Distractor plot for item 6



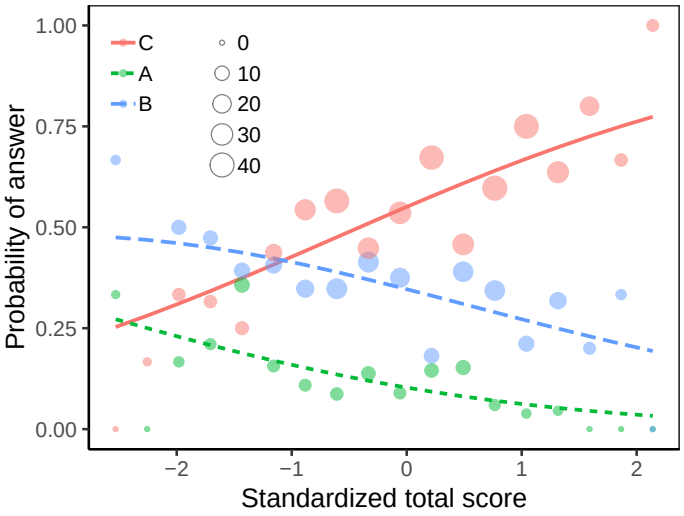
Multinomial plot for item 6



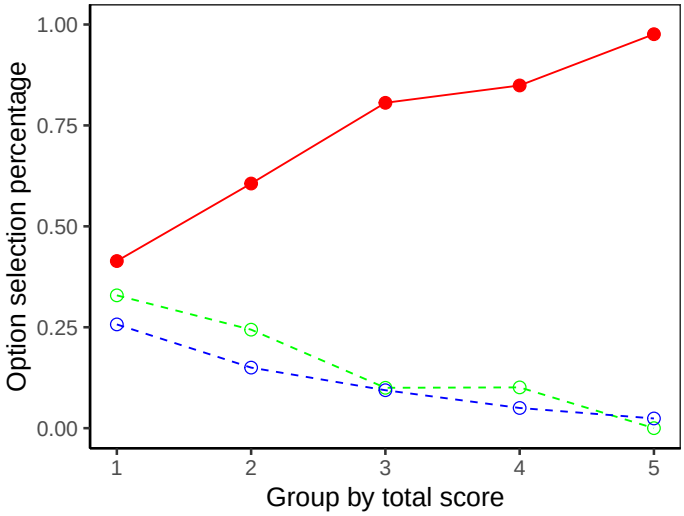
Distractor plot for item 7



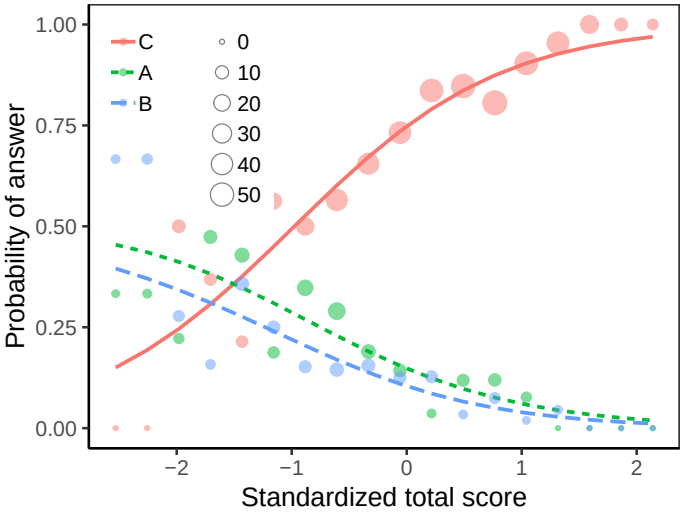
Multinomial plot for item 7



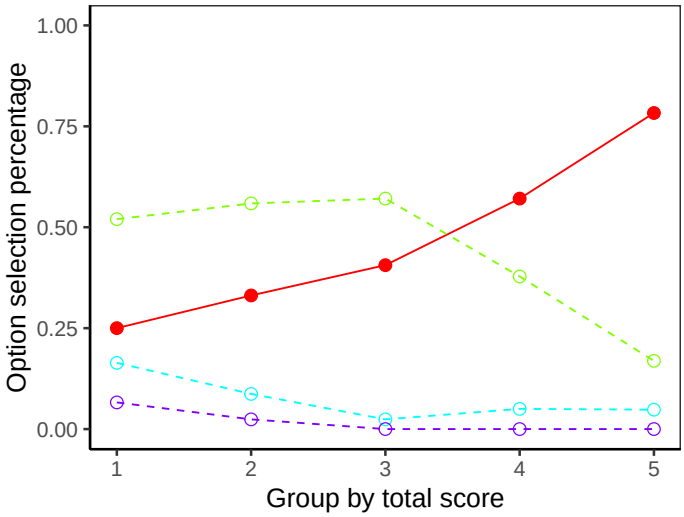
Distractor plot for item 8



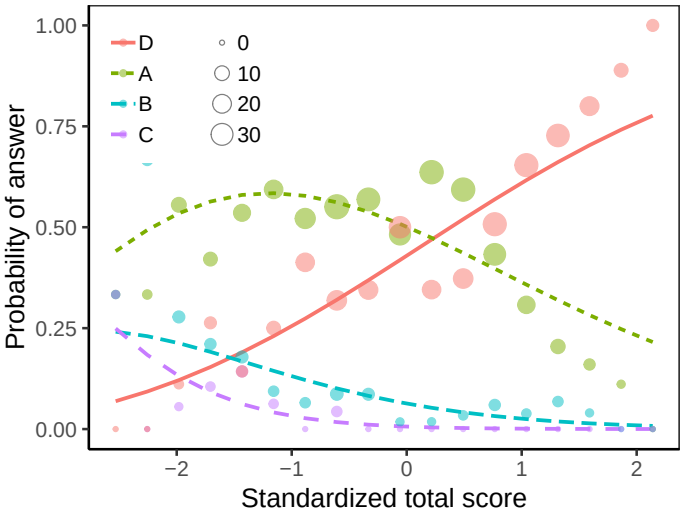
Multinomial plot for item 8



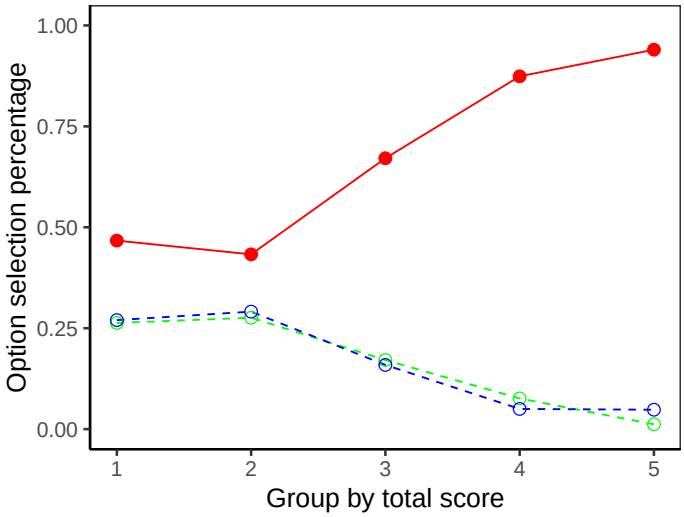
Distractor plot for item 9



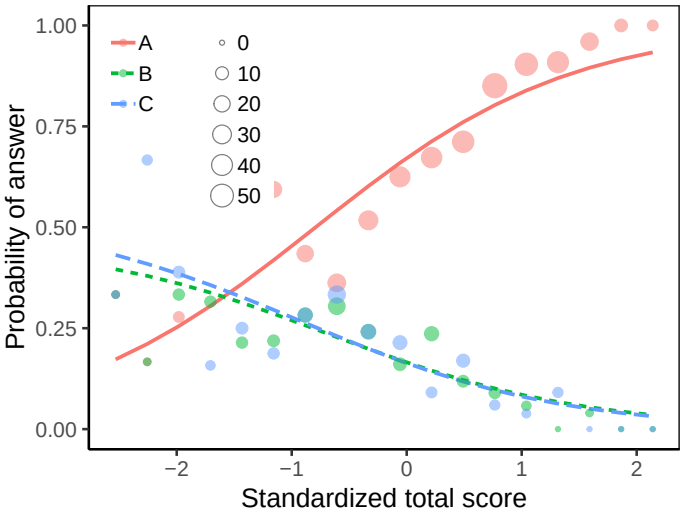
Multinomial plot for item 9



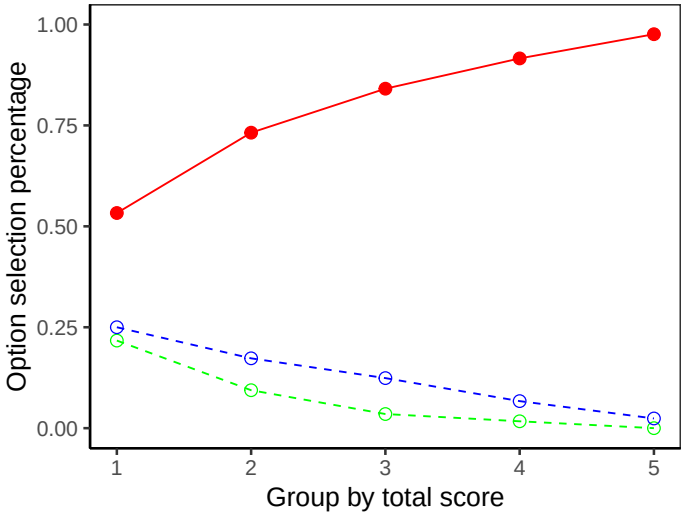
Distractor plot for item 10



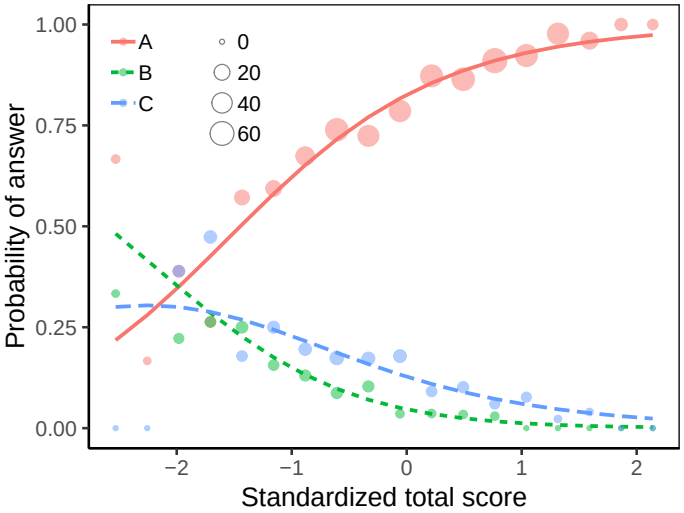
Multinomial plot for item 10



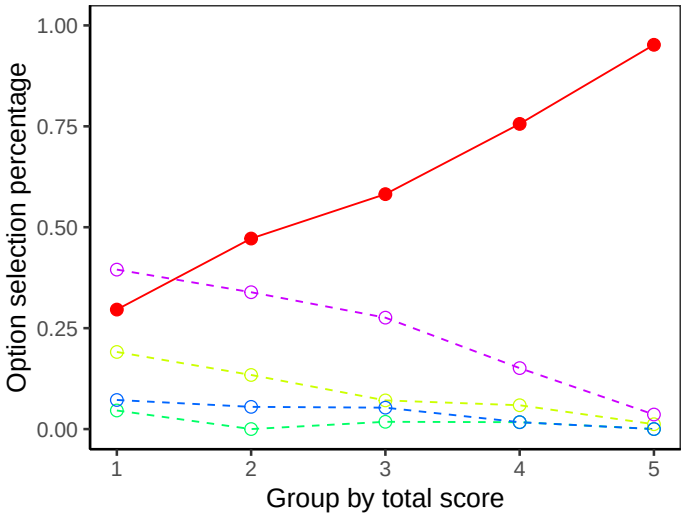
Distractor plot for item 11



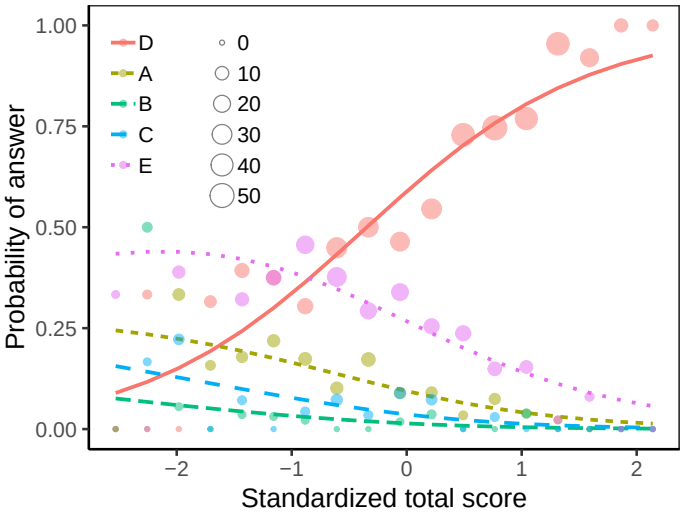
Multinomial plot for item 11



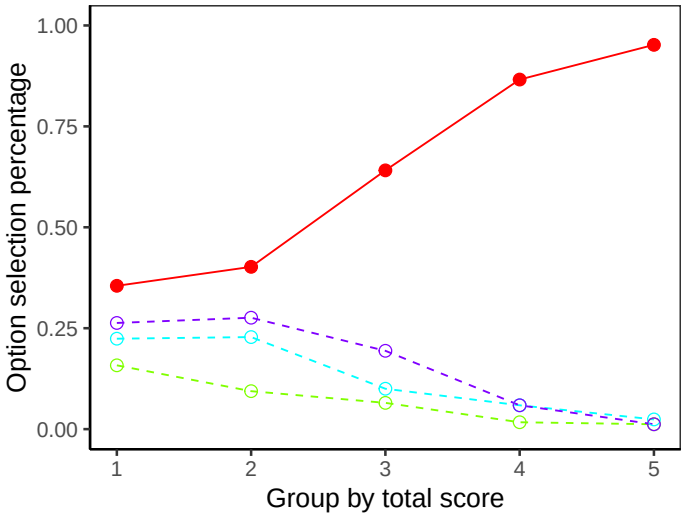
Distractor plot for item 12



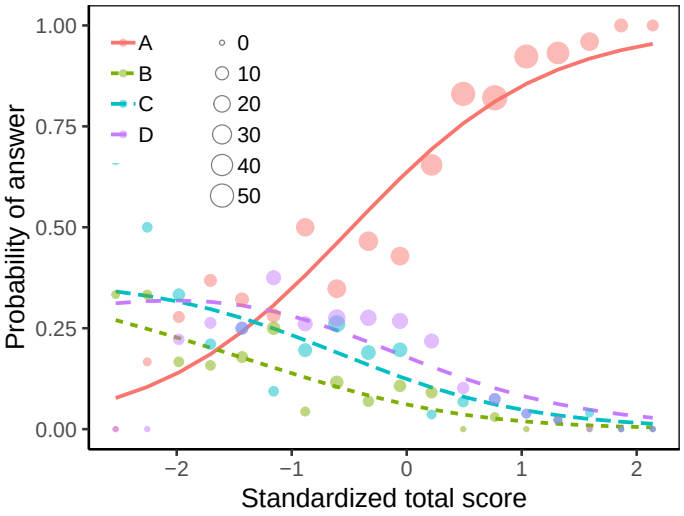
Multinomial plot for item 12



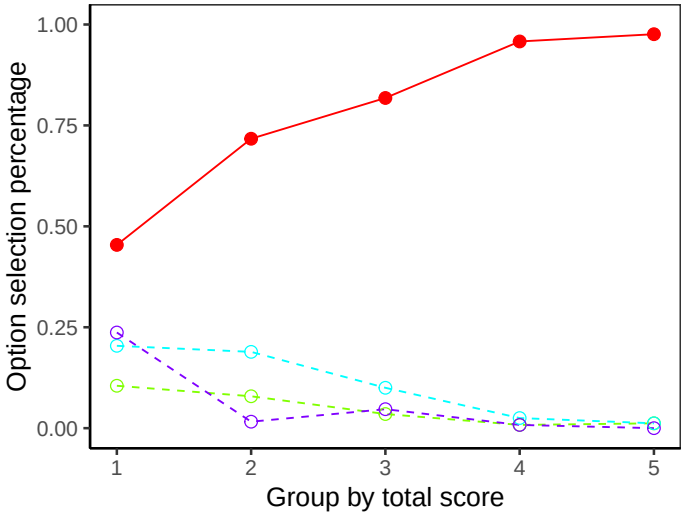
Distractor plot for item 13



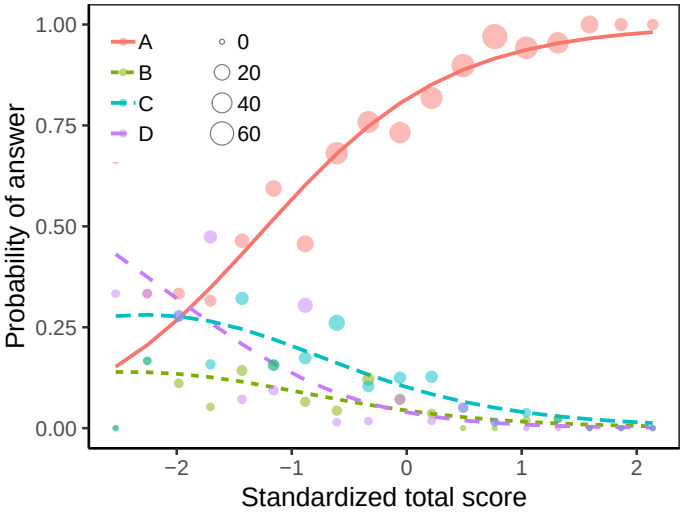
Multinomial plot for item 13



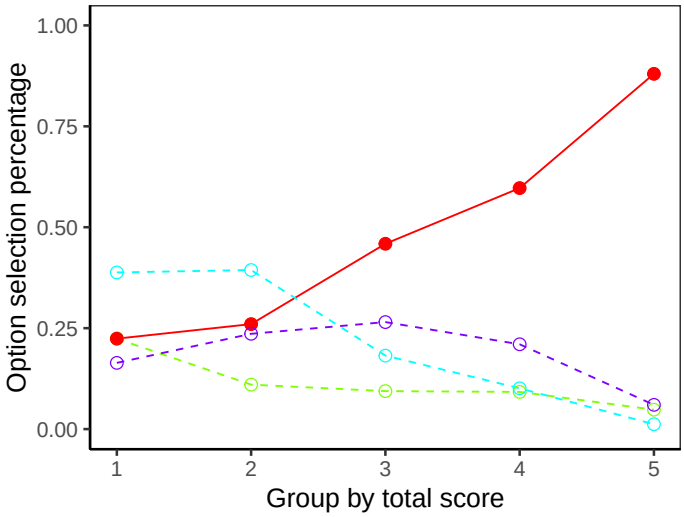
Distractor plot for item 14



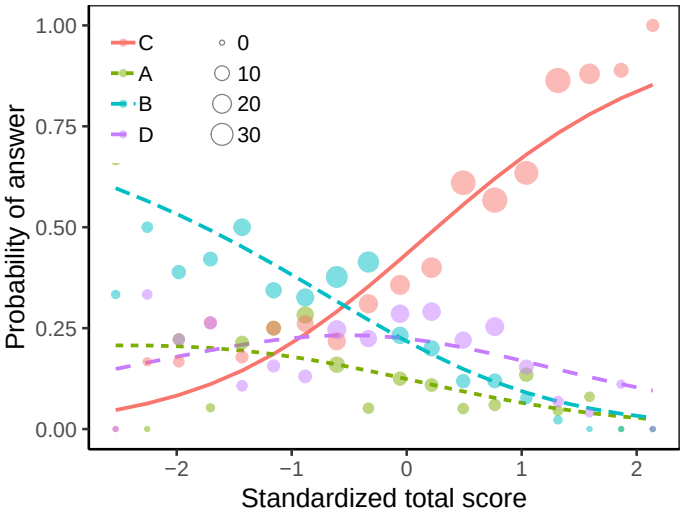
Multinomial plot for item 14



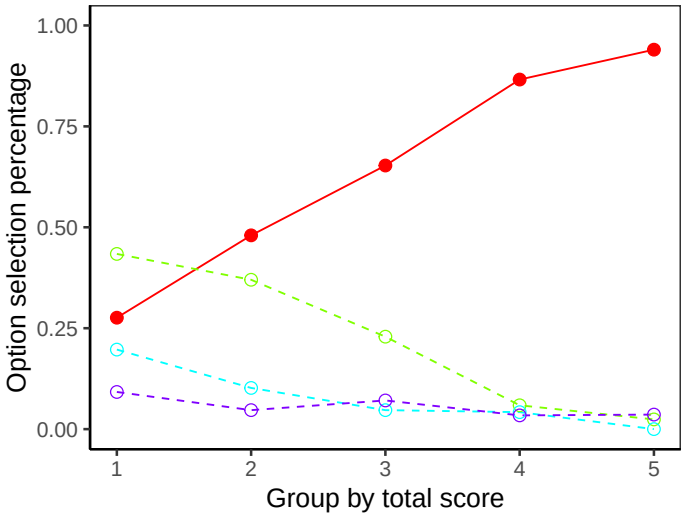
Distractor plot for item 15



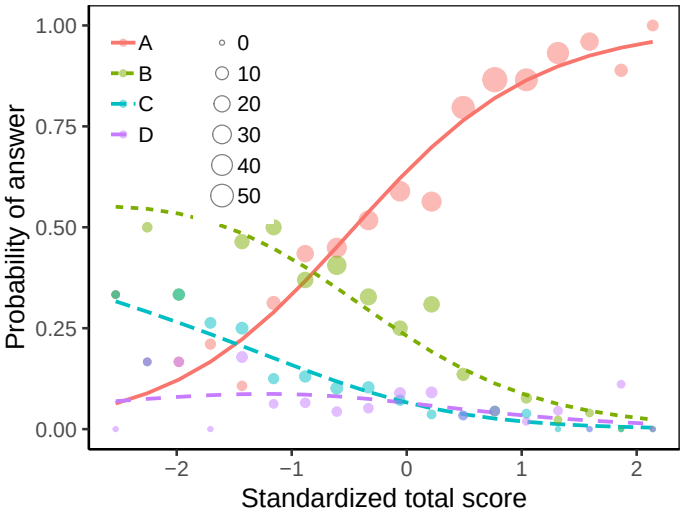
Multinomial plot for item 15



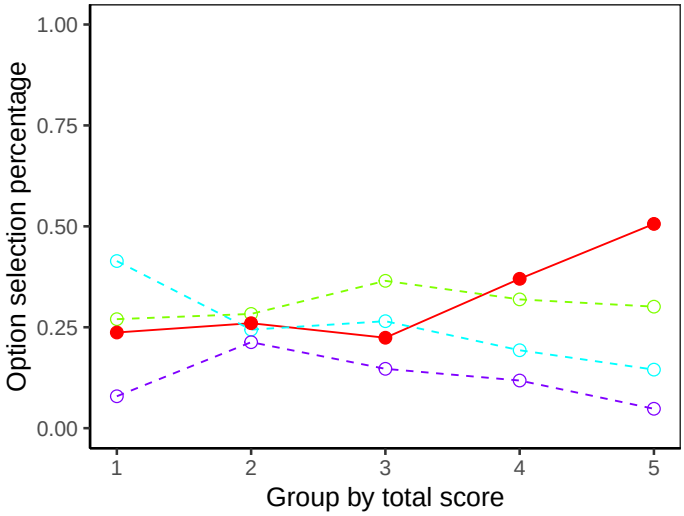
Distractor plot for item 16



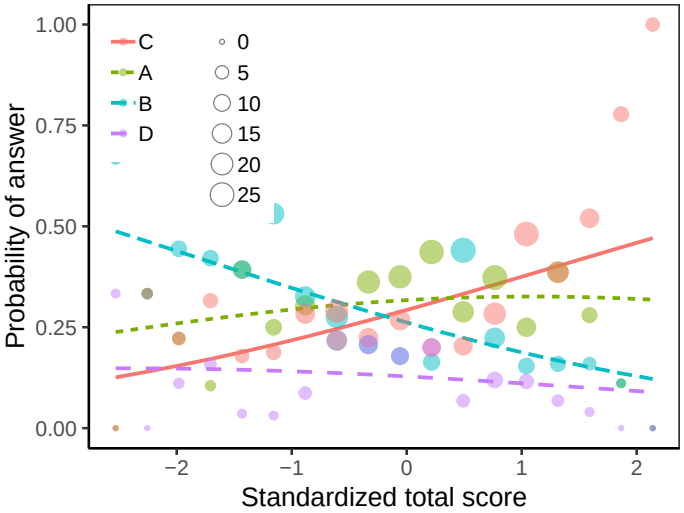
Multinomial plot for item 16



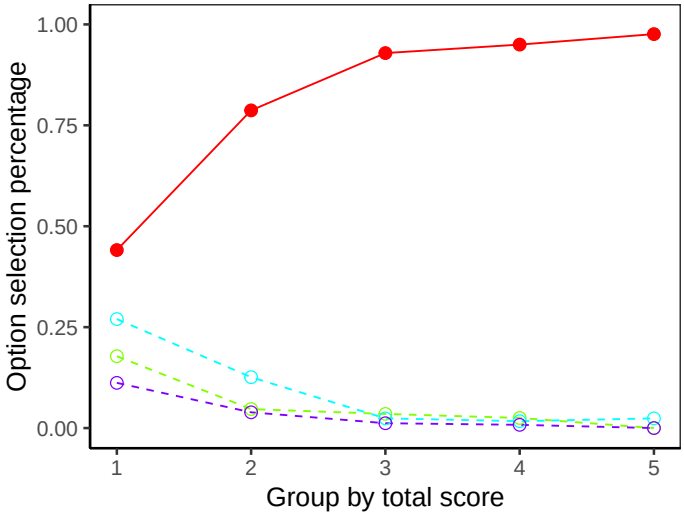
Distractor plot for item 17



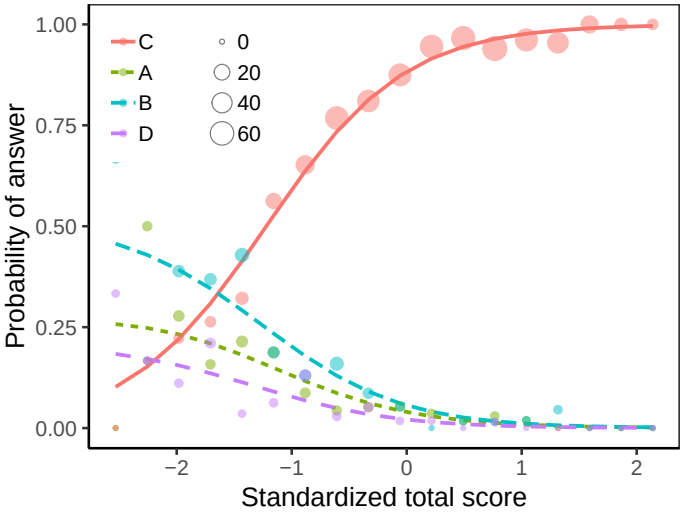
Multinomial plot for item 17



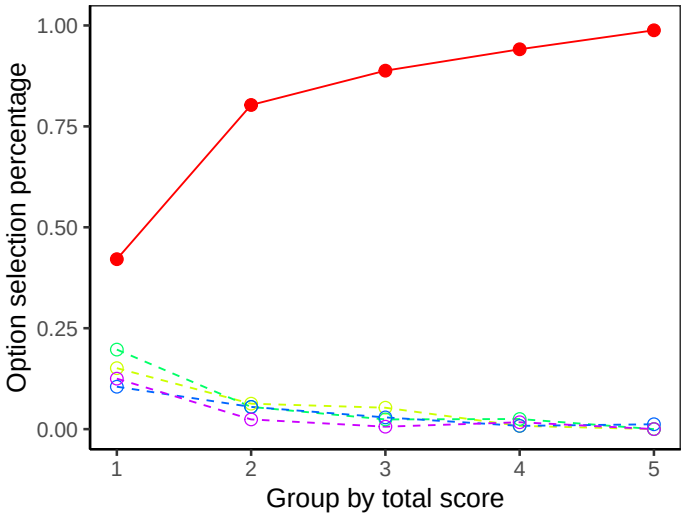
Distractor plot for item 18



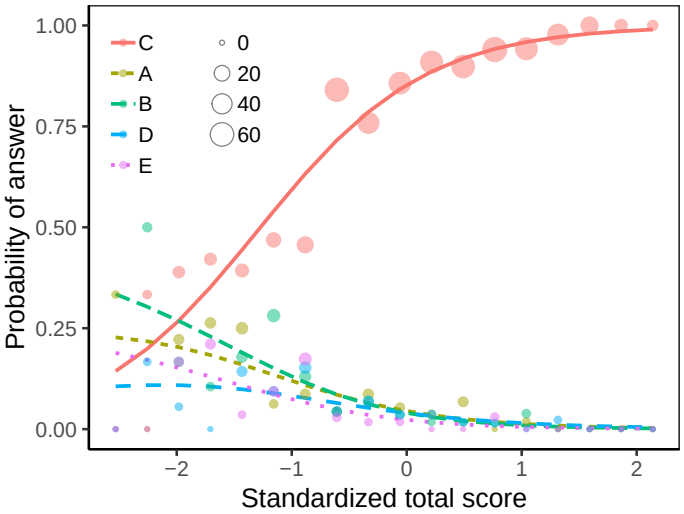
Multinomial plot for item 18



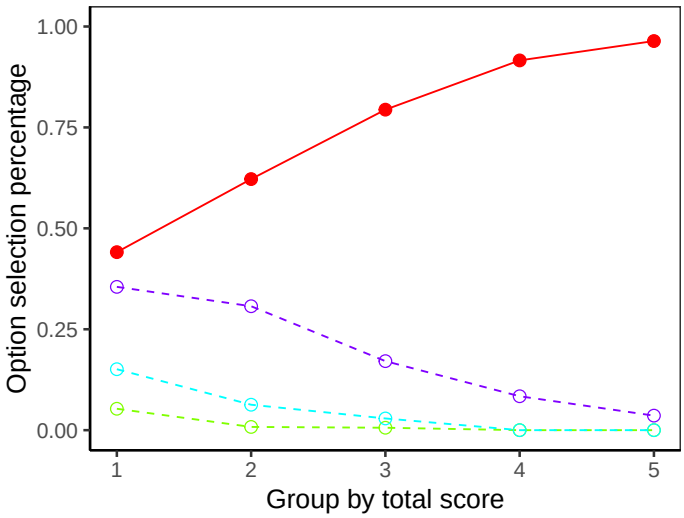
Distractor plot for item 19



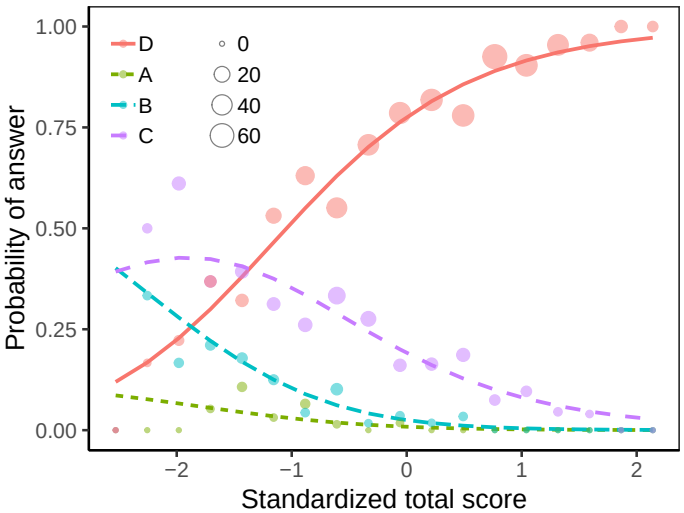
Multinomial plot for item 19



Distractor plot for item 20



Multinomial plot for item 20

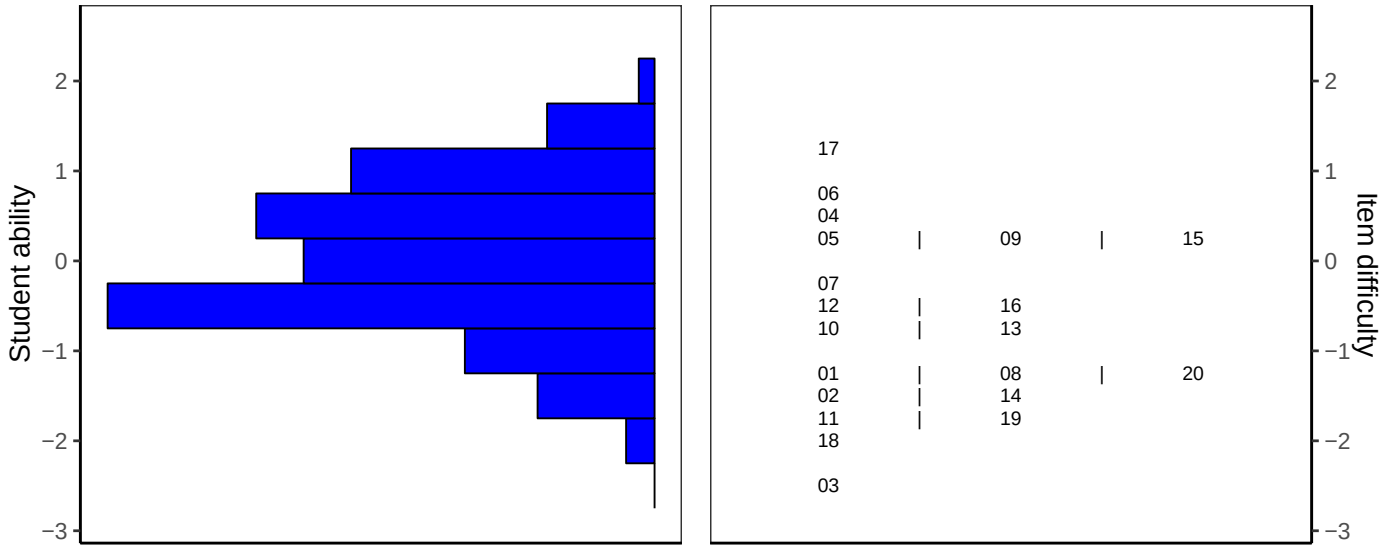


IRT models

Item Response Theory (IRT) models are mixed-effect regression models in which the respondent's ability θ is assumed to be a latent and is estimated together with item parameters.

Wright (item-person) map using 1PL IRT model

A Wright map, also called an item-person map, is a graphical tool to display person estimates and item parameters. The person side (left) represents a histogram of estimated knowledge of the respondents. The item side (right) displays estimates of the difficulty of particular items.



Selected model

All subsequent analyses are based on the selected 3PL IRT model:

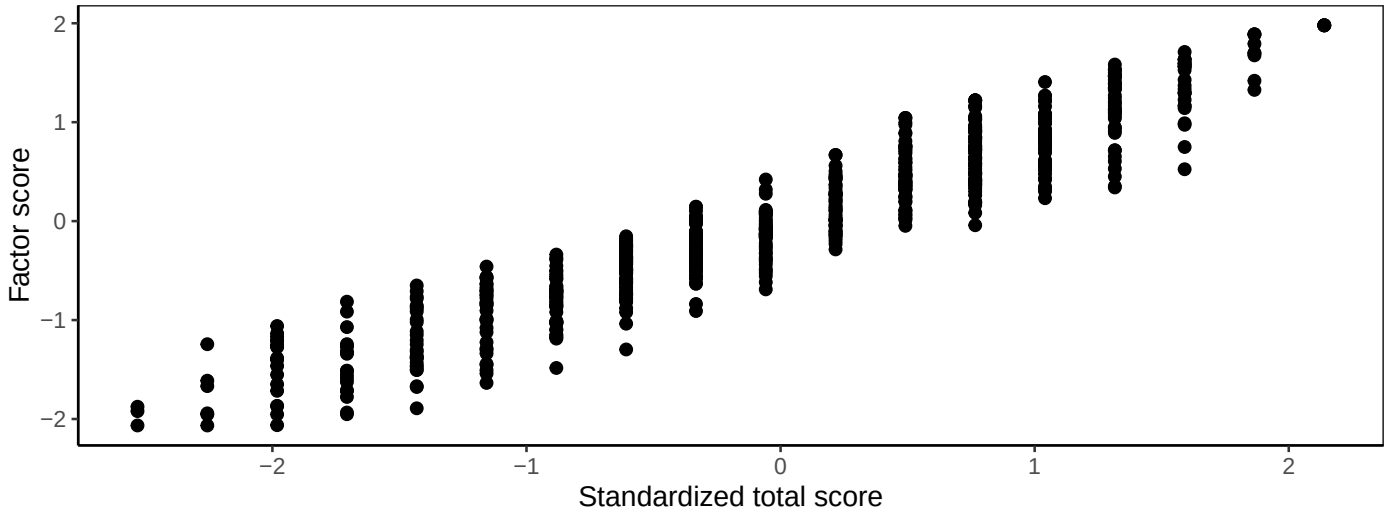
$$P(Y_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + (1 + c_j)\frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$$

Model parameters are estimated using a marginal maximum likelihood method. Ability θ is assumed to follow standard normal distribution.

Ability estimates

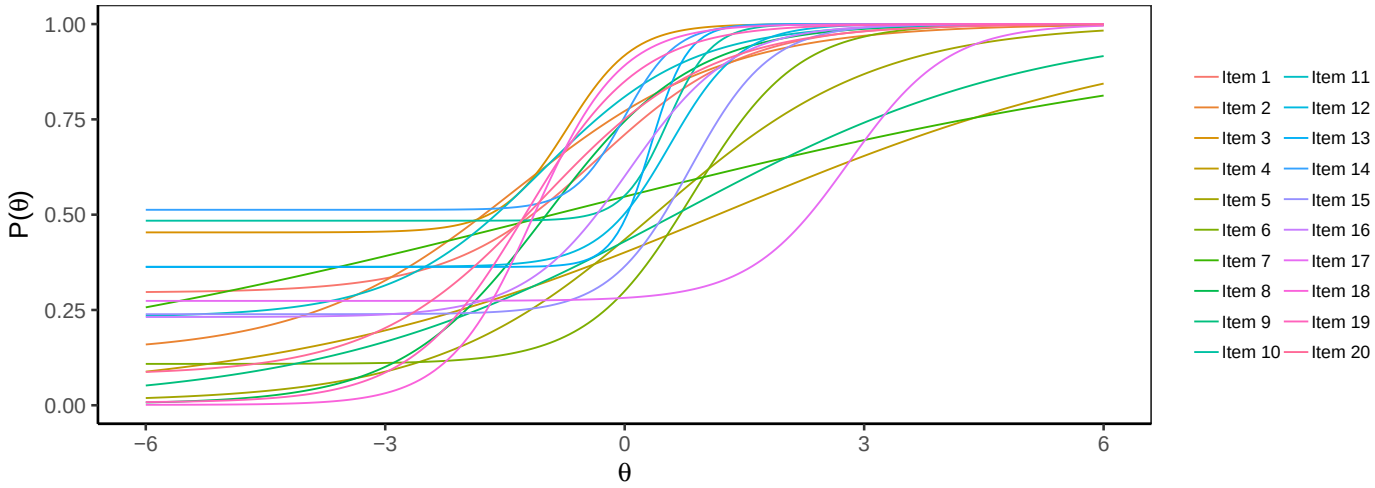
Ability is estimated using an EAP algorithm and a selected IRT model. The relationship between ability estimates (factor scores, F-scores) and standardized total test scores (Z-scores) can be seen on the scatter plot below. A table with ability estimates for all respondents can be downloaded from the application.

	Min	Max	Mean	Median	SD	Skewness	Kurtosis
Total Scores	3.00	20.00	12.21	12.00	3.64	-0.20	2.35
Z-Scores	-2.53	2.14	0.00	-0.06	1.00	-0.20	2.35
F-scores	-2.07	1.98	-0.00	-0.04	0.88	-0.04	2.43

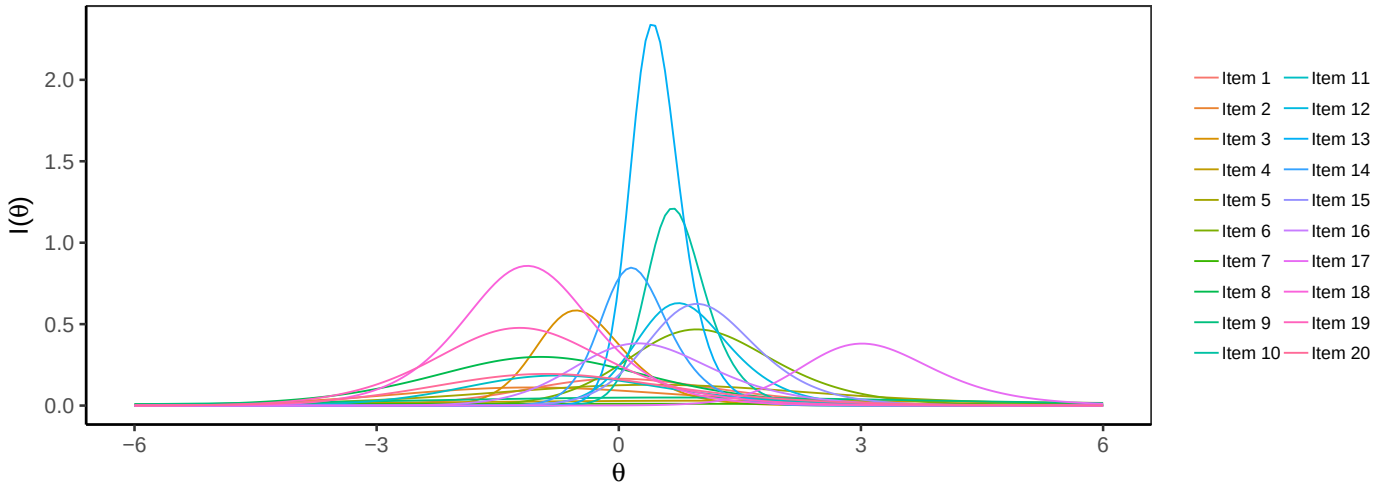


Item characteristic and information curves

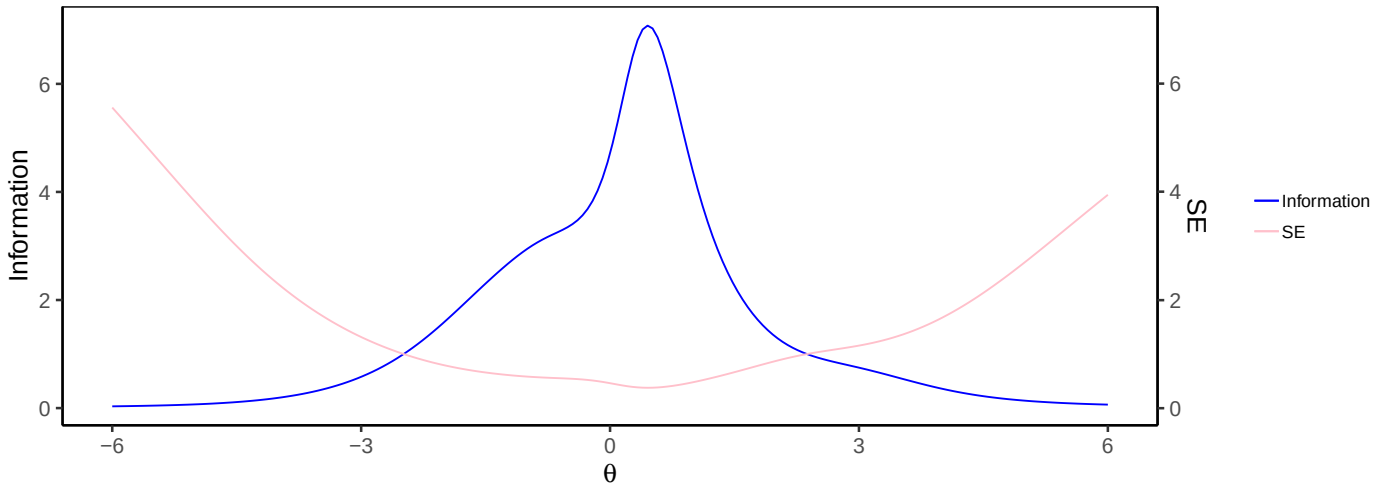
Item characteristic curves



Item information curves



Test information function



Parameter estimates and item fit

Estimates of parameters are completed by their standard errors (SE) and by signed Chi-squared statistics SX2 (see Orlando and Thissen, 2000). P-values lower than 0.05 indicate suspicious items are not fitting the selected IRT model. SX2 is computed only when no missing data are present.

	a	SE(a)	b	SE(b)	c	SE(c)	d	SE(d)	SX2-value	df	p-value
Item 1	1.08	0.37	-0.33	0.68	0.30	1.03	1.00	-	13.26	12	0.35
Item 2	0.76	0.36	-1.35	2.20	0.14	6.17	1.00	-	18.81	12	0.09
Item 3	2.41	0.92	-0.72	0.41	0.45	0.70	1.00	-	13.66	9	0.14
Item 4	0.35	0.15	1.27	1.23	0.02	11.93	1.00	-	15.41	13	0.28
Item 5	0.72	0.17	0.39	0.38	0.01	12.59	1.00	-	13.03	11	0.29
Item 6	1.52	0.43	0.85	0.14	0.11	0.62	1.00	-	14.64	11	0.20
Item 7	0.21	0.10	-0.75	2.00	0.02	13.40	1.00	-	21.30	13	0.07
Item 8	1.10	0.14	-0.97	0.15	0.00	11.89	1.00	-	19.62	11	0.05
Item 9	0.45	0.10	0.65	0.28	0.00	11.87	1.00	-	16.45	13	0.23
Item 10	3.60	1.44	0.54	0.12	0.48	0.16	1.00	-	15.92	11	0.14
Item 11	1.07	0.42	-1.03	1.21	0.23	2.57	1.00	-	4.74	11	0.94
Item 12	2.26	0.75	0.56	0.16	0.36	0.25	1.00	-	14.57	12	0.27
Item 13	4.37	1.83	0.33	0.10	0.36	0.19	1.00	-	16.90	11	0.11
Item 14	3.11	1.22	-0.00	0.20	0.51	0.28	1.00	-	12.22	10	0.27
Item 15	1.99	0.58	0.82	0.13	0.24	0.26	1.00	-	8.81	12	0.72
Item 16	1.54	0.38	0.05	0.24	0.23	0.55	1.00	-	10.97	12	0.53
Item 17	1.60	1.47	2.81	1.03	0.27	0.15	1.00	-	15.04	13	0.30
Item 18	1.85	0.24	-1.14	0.11	0.00	11.95	1.00	-	9.26	10	0.51
Item 19	1.39	0.19	-1.24	0.19	0.01	11.95	1.00	-	16.32	11	0.13
Item 20	0.95	0.31	-1.04	1.03	0.08	5.70	1.00	-	9.06	11	0.62

DIF/Fairness analysis

Total scores by group

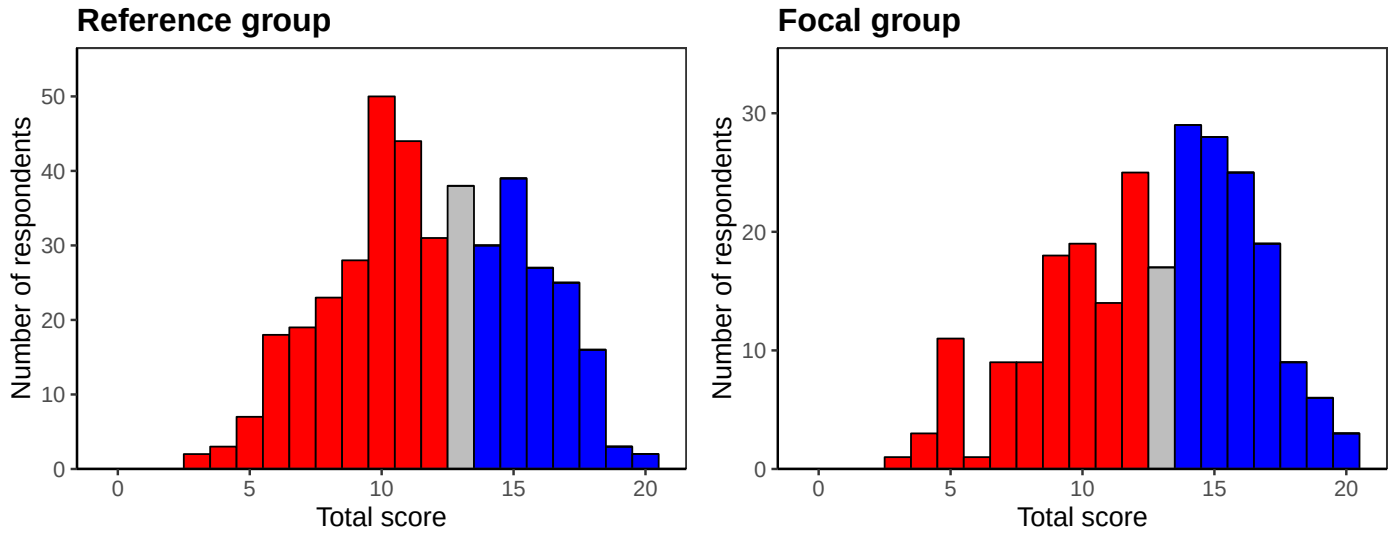
Summary table of total scores in a reference and focal group

DIF is not about total scores! Two groups may have the same distribution of total scores, yet, some item may function differently for the two groups. Also, one of the groups may have a signifficantly lower total score, yet, it may happen that there is no DIF item!

	Min	Max	Mean	Median	SD	Skewness	Kurtosis
Reference group (0)	3	20	11.92	12.00	3.55	-0.08	2.31
Focal group (1)	3	20	12.70	13.00	3.74	-0.42	2.53

Histogram of total score by group

For a selected cut-score, the blue part of histograms shows respondents with a total score above the cut-score, the grey column shows respondents with a total score equal to the cut-score and the red part of histograms shows respondents below the cut-score.



Delta plot method

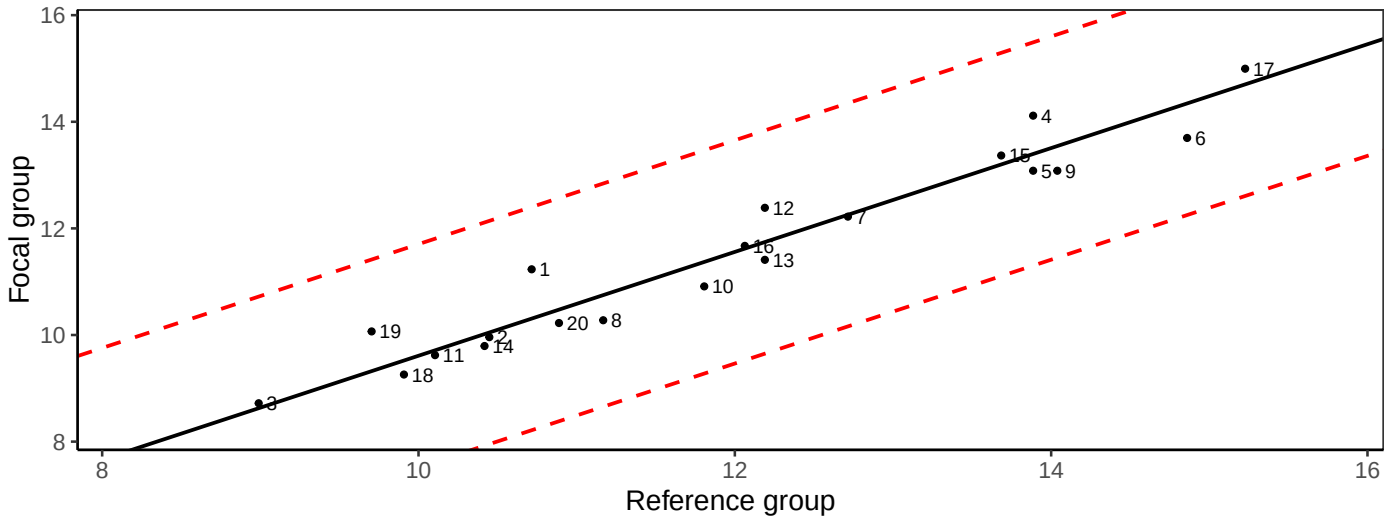
A delta plot compares the proportions of correct answers per item in the two groups. It displays non-linear transformation of these proportions using quantiles of standard normal distributions (so called delta scores) of each item for the two groups in a scatterplot called diagonal plot or delta plot . An item is under suspicion of DIF if the delta point departs considerably from the major axis.

Summary table

Detection threshold is 1.5. **No DIF item was detected.**
Item purification was not applied.

	Prop. Ref	Prop. Foc	Delta Ref	Delta Foc	Dist.
Item 1	0.716	0.671	10.715	11.232	-0.662
Item 2	0.738	0.776	10.448	9.959	0.063
Item 3	0.842	0.858	8.990	8.719	-0.067
Item 4	0.412	0.390	13.886	14.115	-0.515
Item 5	0.412	0.492	13.886	13.082	0.225
Item 6	0.321	0.431	14.860	13.696	0.464
Item 7	0.528	0.577	12.715	12.221	0.025
Item 8	0.677	0.752	11.168	10.276	0.338
Item 9	0.398	0.492	14.039	13.082	0.332
Item 10	0.617	0.699	11.807	10.912	0.328
Item 11	0.765	0.801	10.104	9.622	0.065
Item 12	0.580	0.561	12.190	12.386	-0.460
Item 13	0.580	0.654	12.190	11.410	0.239
Item 14	0.741	0.789	10.417	9.793	0.160
Item 15	0.432	0.463	13.684	13.367	-0.121
Item 16	0.593	0.630	12.063	11.672	-0.037
Item 17	0.289	0.309	15.227	14.995	-0.211
Item 18	0.780	0.825	9.908	9.258	0.188
Item 19	0.795	0.768	9.704	10.067	-0.534
Item 20	0.701	0.756	10.888	10.225	0.180

Delta plot


Focal group
Reference group
16
14
12
10
8
8
10
12
14
16
17
4
6
15
5
9
12
7
16
13
1
10
19
20
8
2
14
11
18
3

Logistic regression method

Logistic regression allows for detection of uniform and non-uniform DIF by adding a group specific intercept (uniform DIF) and group specific interaction (non-uniform DIF) into the model and by testing for their significance.

Summary table

Both types of DIF were tested. The detection threshold is 5.99. **No DIF item was detected.**
Item purification was not applied. Benjamini-Hochberg p-value adjustment for multiple comparisons was used.

	Stat.	P-value	R^2	ZT	JG
Item.1	6.543	0.212	0.012	A	A
Item.2	0.119	0.942	0.000	A	A
Item.3	0.401	0.942	0.001	A	A
Item.4	3.029	0.550	0.006	A	A
Item.5	1.358	0.845	0.002	A	A
Item.6	2.688	0.580	0.004	A	A
Item.7	0.380	0.942	0.001	A	A
Item.8	1.716	0.771	0.003	A	A
Item.9	3.333	0.550	0.006	A	A
Item.10	2.427	0.594	0.004	A	A
Item.11	4.996	0.329	0.010	A	A
Item.12	6.318	0.212	0.011	A	A
Item.13	3.229	0.550	0.005	A	A
Item.14	0.255	0.942	0.000	A	A
Item.15	0.427	0.942	0.001	A	A
Item.16	0.207	0.942	0.000	A	A
Item.17	0.545	0.942	0.001	A	A
Item.18	0.269	0.942	0.000	A	A
Item.19	9.703	0.102	0.018	A	A
Item.20	9.178	0.102	0.016	A	A

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Effect size is based on Nagelkerke's R^2.
'A' means negligible, 'B' moderate and 'C' large effect size
The thresholds are:
Zumbo & Thomas (ZT): 0 'A' 0.13 'B' 0.26 'C' 1
Jodoin & Gierl (JG): 0 'A' 0.035 'B' 0.07 'C' 1.

DDF detection using multinomial regression method

Differential Distractor Functioning (DDF) occurs when respondents from different groups but with the same knowledge have a different probability of selecting at least one distractor choice. DDF is examined here by a multinomial log-linear regression model with Z-score and group membership as covariates.

Summary table

Both types were of DDF tested. **No DDF item was detected.**
Item purification was not applied. Benjamini-Hochberg p-value adjustment for multiple comparisons was used.

	Stat.	P-value	Adj. p-value
Item.1	6.886	0.331	0.741
Item.2	3.822	0.431	0.783
Item.3	2.368	0.668	0.836
Item.4	5.579	0.472	0.787
Item.5	2.611	0.856	0.922
Item.6	3.823	0.431	0.783
Item.7	2.701	0.609	0.812
Item.8	1.758	0.780	0.918
Item.9	7.095	0.312	0.741
Item.10	6.386	0.172	0.574
Item.11	6.642	0.156	0.574
Item.12	18.503	0.018	0.178
Item.13	9.103	0.168	0.574
Item.14	2.431	0.876	0.922
Item.15	0.622	0.996	0.996
Item.16	5.104	0.531	0.805
Item.17	4.850	0.563	0.805
Item.18	6.867	0.333	0.741
Item.19	19.942	0.011	0.178
Item.20	12.078	0.060	0.402

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Session info

Session info provides information about settings of the R console and used packages and their versions which were used for the analysis.

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.1 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4    grid      stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] xtable_1.8-3            stringr_1.3.1
##  [3] ShinyItemAnalysis_1.2.9 rmarkdown_1.10
##  [5] psychometric_2.2        multilevel_2.6
##  [7] nlme_3.1-131            psych_1.8.4
##  [9] moments_0.14            mirt_1.29
## [11] ltm_1.1-1               polycor_0.7-9
## [13] latticeExtra_0.6-28     RColorBrewer_1.1-2
## [15] lattice_0.20-35         knitr_1.20
## [17] gridExtra_2.3           ggdendro_0.1-20
## [19] difR_5.0                difNLR_1.2.2
## [21] reshape2_1.4.3          nnet_7.3-12
## [23] msm_1.6.6               data.table_1.11.6
## [25] deltaPlotR_1.6          MASS_7.3-49
## [27] CTT_2.3.3               corrplot_0.84
## [29] shinyjs_1.0             shinydashboard_0.7.0
## [31] shinyBS_0.61            plotly_4.8.0
## [33] ggplot2_3.0.0           DT_0.4
## [35] shiny_1.1.0
##
## loaded via a namespace (and not attached):
##  [1] httr_1.3.1            rprojroot_1.3-2       Deriv_3.8.5
```

```
##  [4] tools_3.4.4           backports_1.1.2       R6_2.2.2
##  [7] vegan_2.5-2           lazyeval_0.2.1        mgcv_1.8-23
## [10] colorspace_1.3-2      permute_0.9-4         withr_2.1.2
## [13] tidyselect_0.2.4      mnormt_1.5-5          compiler_3.4.4
## [16] expm_0.999-2          labeling_0.3          scales_1.0.0
## [19] mvtnorm_1.0-8         digest_0.6.17         foreign_0.8-69
## [22] minqa_1.2.4           pkgconfig_2.0.2       htmltools_0.3.6
## [25] lme4_1.1-18-1         htmlwidgets_1.2       rlang_0.2.2
## [28] bindr_0.1.1           jsonlite_1.5          crosstalk_1.0.0
## [31] dcurver_0.9.1         dplyr_0.7.6           magrittr_1.5
## [34] Matrix_1.2-12         Rcpp_0.12.18          munsell_0.5.0
## [37] stringi_1.2.4         yaml_2.2.0            plyr_1.8.4
## [40] parallel_3.4.4        promises_1.0.1        crayon_1.3.4
## [43] cowplot_0.9.3         splines_3.4.4         pillar_1.3.0
## [46] GPArotation_2014.11-1 glue_1.3.0            evaluate_0.11
## [49] nloptr_1.0.4          httpuv_1.4.5          gtable_0.2.0
## [52] purrr_0.2.5           tidyr_0.8.1           assertthat_0.2.0
## [55] xfun_0.3              mime_0.5              later_0.7.5
## [58] survival_2.41-3       viridisLite_0.3.0     tibble_1.4.2
## [61] tinytex_0.8           bindrcpp_0.2.2        cluster_2.0.6
```