Certification exam, Form A: Technical report

Psychometrics team (fictional example) | November 03, 2026 | mirt 1.47 | mirtverse 0.2.0

Contents

1. Summary
2. Model
3. Items
4. Graphics
5. Scores
6. Item health cards
7. Blueprint
8. Field-test design
9. Groups
10. Cut scores
11. Scoring
12. Test security
13. Pipeline decisions
14. Reproducibility

Summary

Overview

- A 2PL model with 1 dimension was fitted to 1200 respondents and 31 items (EM estimation).
- The estimation converged normally.

Model fit

- Overall model fit is close: M2 = 561.36, df = 434, RMSEA = 0.016, CFI = 0.987, SRMSR = 0.030.
- The chi-square test is significant, but with large samples this is common; the RMSEA indicates the misfit is small in practical terms.

Reliability

- Marginal reliability is 0.83 (good).

Targeting

- The average item location is -0.18 and the average person score is -0.00, so the test is well targeted for this group.
- Measurement is most precise around theta = -0.52 (information = 6.04, SE = 0.41).

Items

- Low discrimination: Q12 (a = 0.42), Q30 (a = 0.43), Q32 (a = 0.47). These items barely separate high and low scorers; review the key and wording.
- Extreme location: Q32 (b = 3.05). These items are far too easy or too hard for most respondents.
- Low item-total correlation: Q12 (r = 0.16), Q21 (r = 0.19), Q30 (r = 0.19), Q32 (r = 0.17). These items relate weakly to the rest of the test.
- Misfit: Q04 (S-X2 p = .004), Q12 (S-X2 p = .002), Q17 (S-X2 p = .002). The model does not describe responses to these items well (alpha = 0.01); inspect them with gg_empirical() or item_card().

Local dependence

- 1 item pair shows possible local dependence (Q3 exceeds the average of -0.03 by more than 0.2).
- Strongest: Q21 & Q22 (Q3 = 0.27).
- Check whether these items share a stimulus, wording, or content; consider a testlet or bifactor model.

Person fit

- 1.8% of respondents have unusual response patterns (Zh < -1.96), which is within what chance alone would produce.

Model

Statistic	Value
nobs	1200
nitems	31
nfact	1
itemtype	2PL
npar	62
logLik	-20593.4
AIC	41310.7
BIC	41626.3

Statistic	Value
SABIC	41429.4
converged	yes
method	EM
fit_stat	M2
M2	561.4
df	434
p.value	0.000
RMSEA	0.016
SRMSR	0.030
CFI	0.987
TLI	0.986
reliability	0.831

Items

Status uses the item health rules: red for misfit or low discrimination, amber for other flags, green otherwise.

item	status	itemtype	N	mean	item_total_r	S_X2	df	RMSEA	p.value
Q01	GREEN	2PL	1200	0.365	0.319	14.433	22	0	0.885
Q02	GREEN	2PL	1200	0.567	0.375	25.562	21	0.013	0.224
Q03	GREEN	2PL	1200	0.312	0.377	22.891	21	0.009	0.350
Q04	RED	2PL	1200	0.338	0.311	43.356	22	0.028	0.004
Q05	GREEN	2PL	1200	0.629	0.360	20.135	20	0.002	0.450
Q06	GREEN	2PL	1200	0.266	0.301	34.593	21	0.023	0.031
Q07	GREEN	2PL	1200	0.747	0.316	23.277	20	0.012	0.275
Q08	GREEN	2PL	1200	0.347	0.296	16.305	22	0	0.800

item	status	itemtype	N	mean	item_total_r	S_X2	df	RMSEA	p.value
Q09	GREEN	2PL	1200	0.795	0.209	28.219	22	0.015	0.169
Q10	GREEN	2PL	1200	0.499	0.300	31.586	22	0.019	0.085
Q11	GREEN	2PL	1200	0.917	0.281	21.980	16	0.018	0.144
Q12	RED	2PL	1200	0.263	0.157	46.575	22	0.031	0.002
Q13	GREEN	2PL	1200	0.665	0.361	30.598	20	0.021	0.061
Q15	GREEN	2PL	1200	0.759	0.289	26.694	21	0.015	0.181
Q16	GREEN	2PL	1200	0.369	0.342	30.285	22	0.018	0.112
Q17	RED	2PL	1200	0.921	0.233	38.053	17	0.032	0.002
Q18	GREEN	2PL	1200	0.398	0.484	32.216	20	0.023	0.041
Q19	GREEN	2PL	1200	0.484	0.336	15.923	21	0	0.774
Q20	GREEN	2PL	1200	0.690	0.436	24.100	18	0.017	0.152
Q21	AMBER	2PL	1200	0.769	0.191	23.694	23	0.005	0.421
Q22	GREEN	2PL	1200	0.815	0.244	16.518	21	0	0.740
Q23	GREEN	2PL	1200	0.341	0.281	16.358	22	0	0.798
Q24	GREEN	2PL	1200	0.632	0.393	17.984	20	0	0.588
Q25	GREEN	2PL	1200	0.390	0.303	22.183	22	0.003	0.449
Q26	GREEN	2PL	1200	0.427	0.482	24.095	19	0.015	0.193
Q27	GREEN	2PL	1200	0.752	0.292	16.813	21	0	0.722
Q28	GREEN	2PL	1200	0.549	0.331	17.983	21	0	0.650
Q29	GREEN	2PL	1200	0.318	0.373	26.799	21	0.015	0.178
Q30	RED	2PL	1200	0.532	0.186	35.340	23	0.021	0.048
Q31	GREEN	2PL	1200	0.767	0.410	11.500	17	0	0.829
Q32	RED	2PL	1200	0.203	0.167	20.112	23	0	0.635

Item parameters

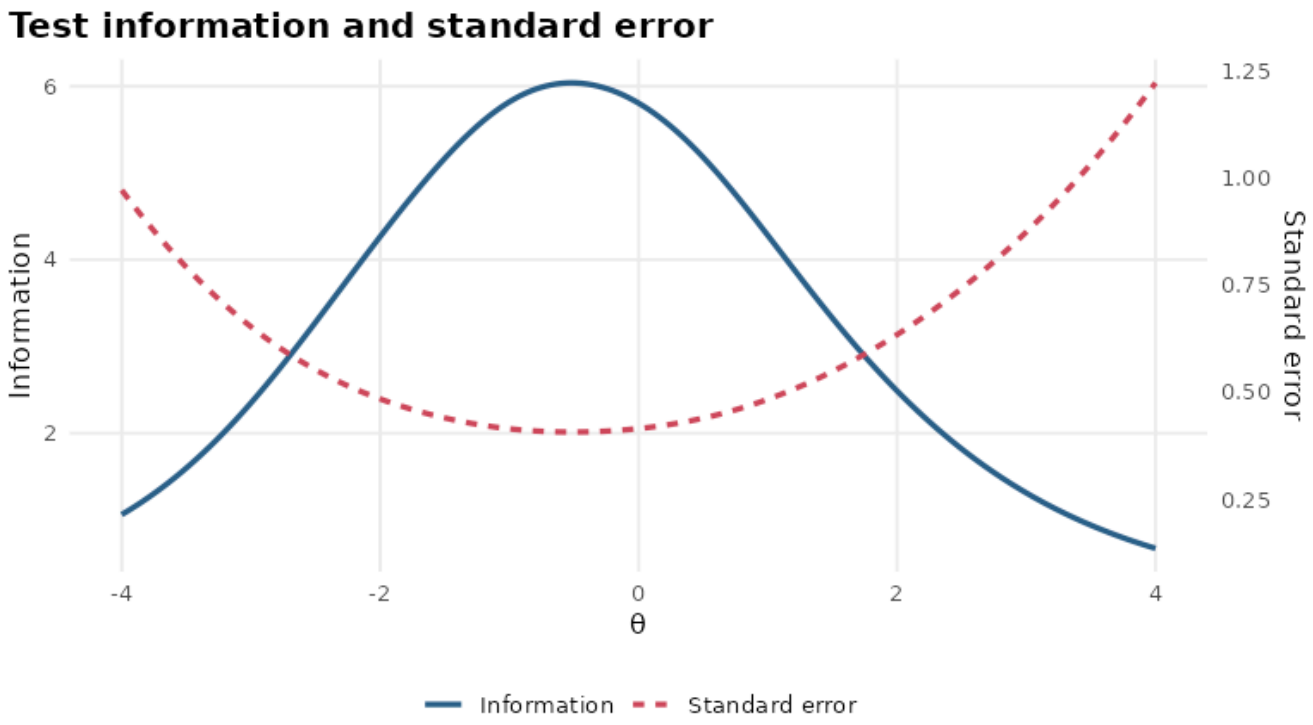
Parameterization: IRT (a, b, g, u).

item	a	b
Q01	0.833	0.750
Q02	1.039	-0.335
Q03	1.036	0.909
Q04	0.770	0.978
Q05	1.059	-0.630
Q06	0.813	1.413
Q07	0.998	-1.301
Q08	0.761	0.926
Q09	0.650	-2.266
Q10	0.749	-0.008
Q11	1.573	-2.046
Q12	0.416	2.568
Q13	1.085	-0.796
Q15	0.901	-1.485
Q16	0.918	0.671
Q17	1.273	-2.373
Q18	1.492	0.357
Q19	0.883	0.067
Q20	1.576	-0.748
Q21	0.542	-2.359
Q22	0.790	-2.103
Q23	0.709	1.023
Q24	1.209	-0.597
Q25	0.793	0.628
Q26	1.519	0.240
Q27	0.914	-1.419

item	a	b
Q28	0.886	-0.277
Q29	1.014	0.896
Q30	0.425	-0.324
Q31	1.652	-1.059
Q32	0.471	3.047

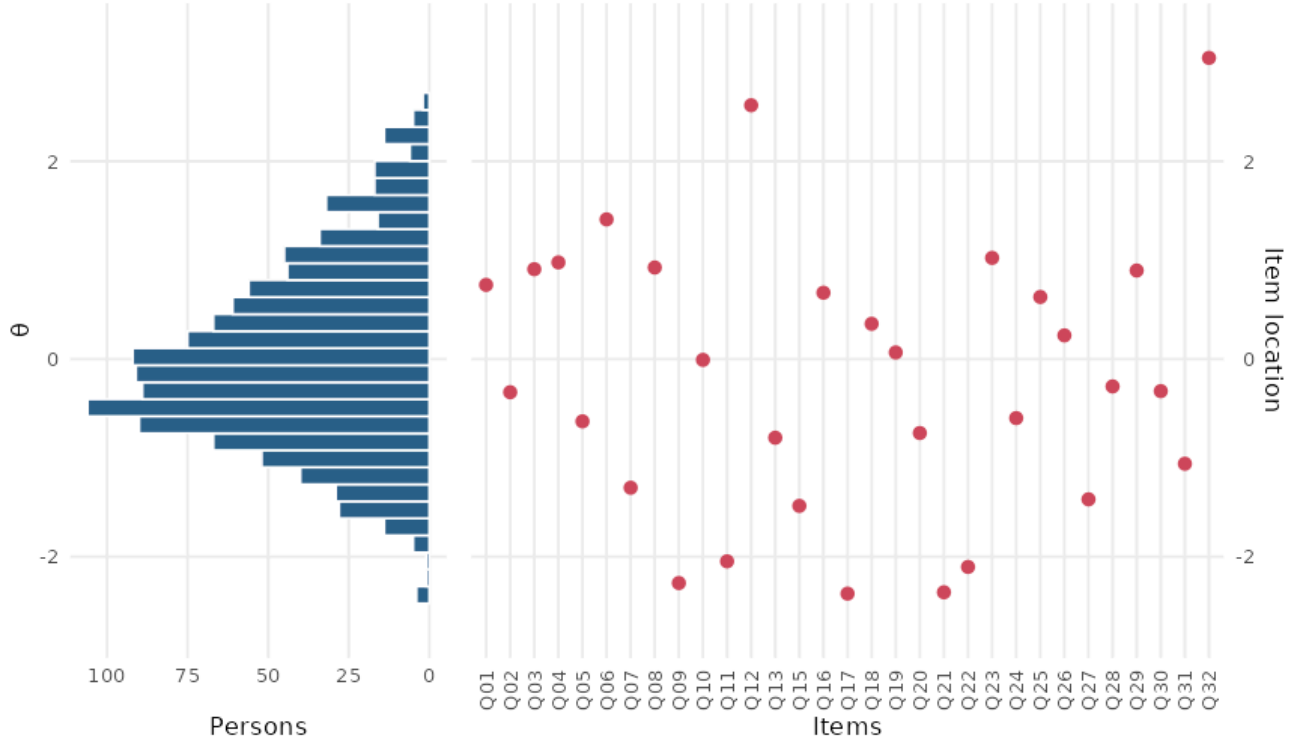
Graphics

Test information



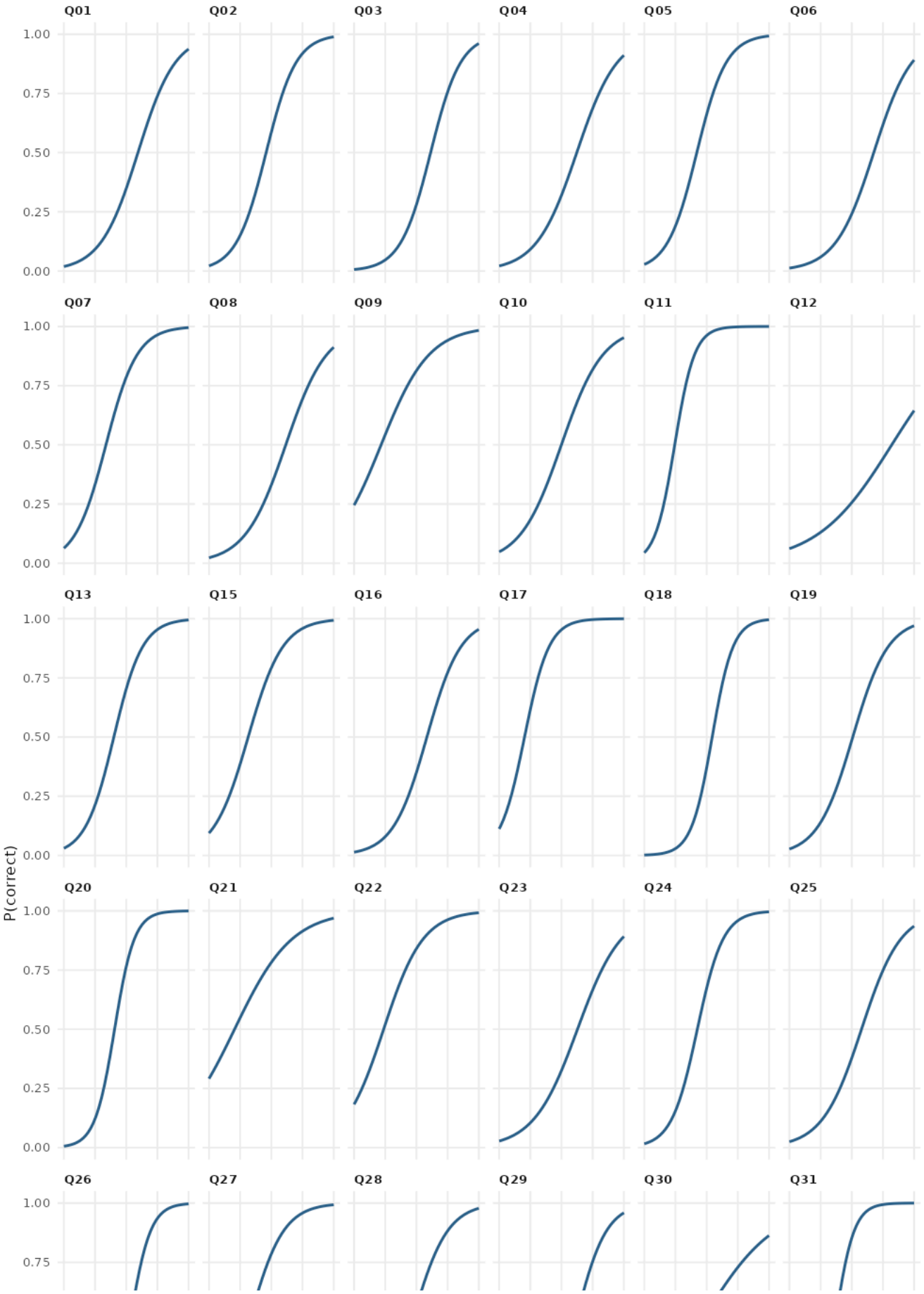
Wright map

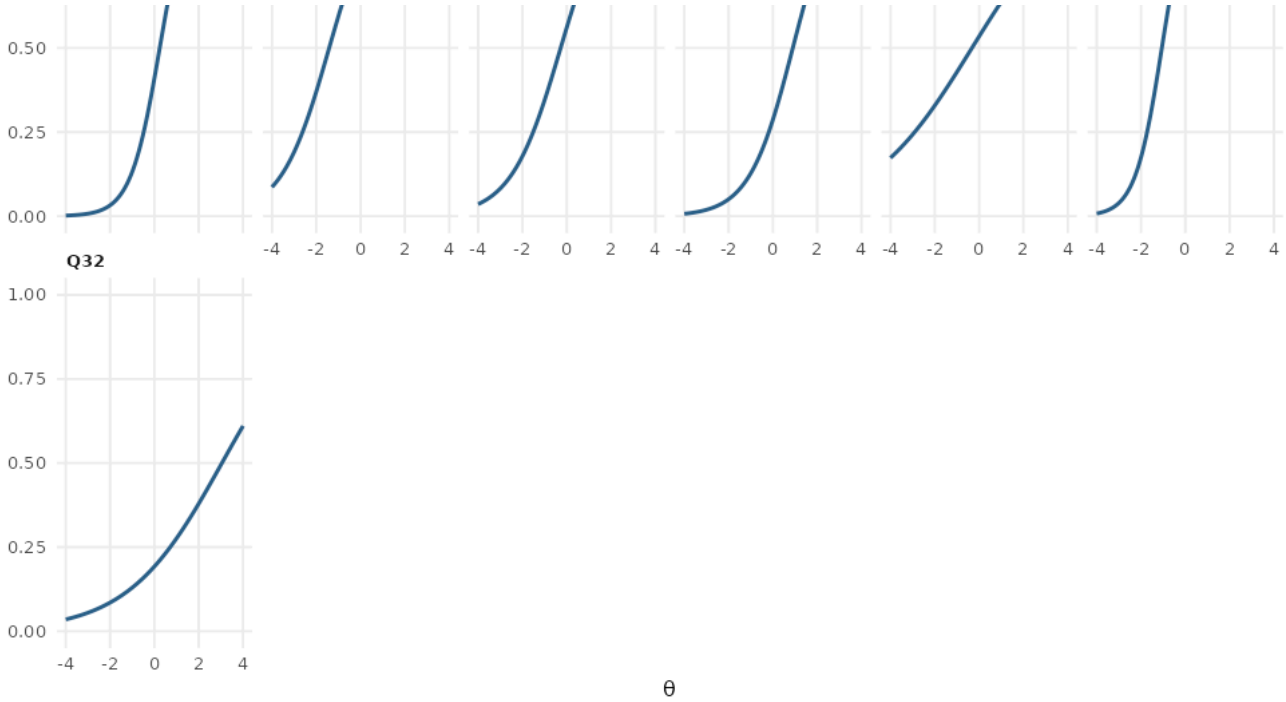
Wright map



Item characteristic curves

Item characteristic curves

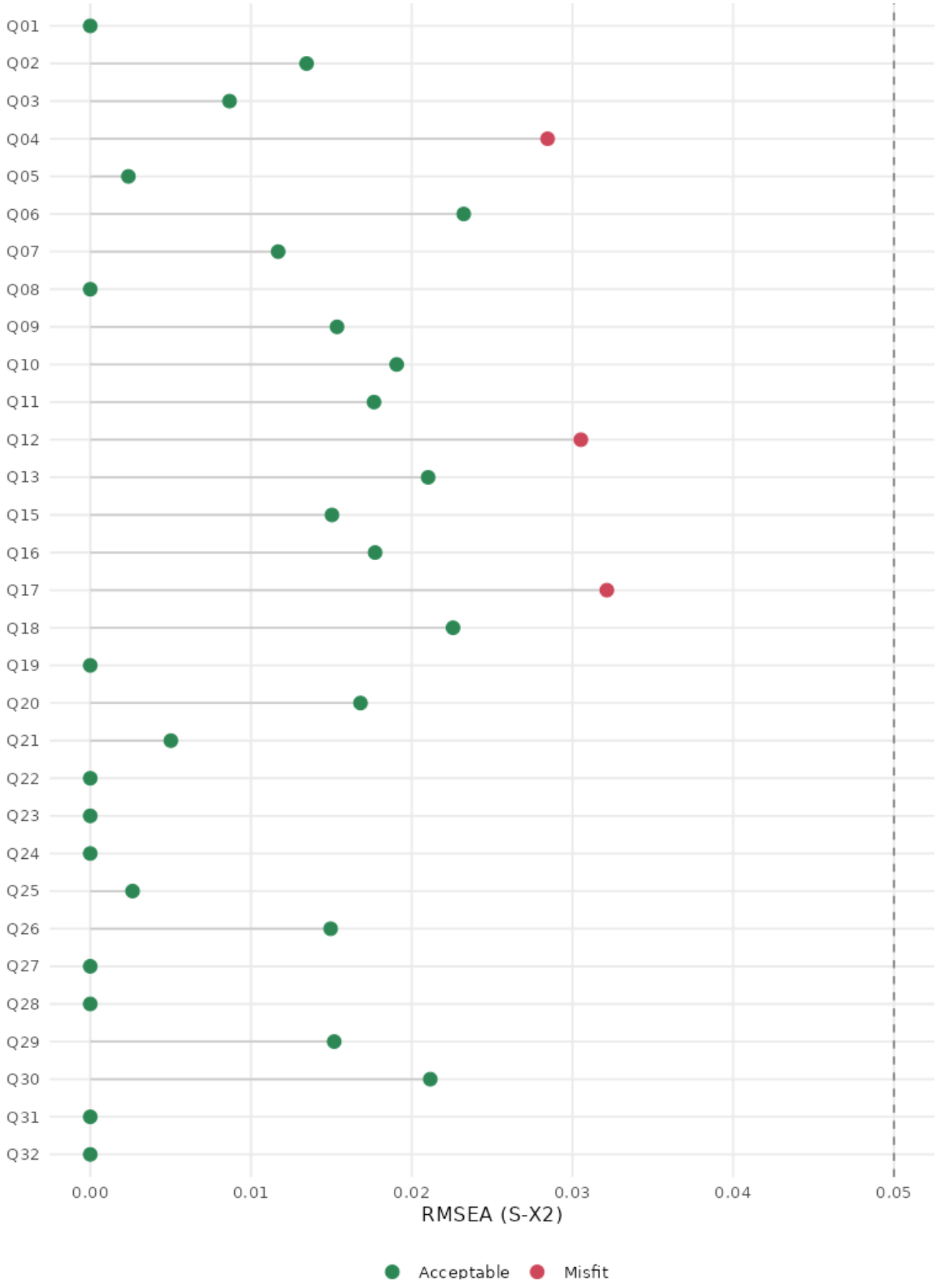




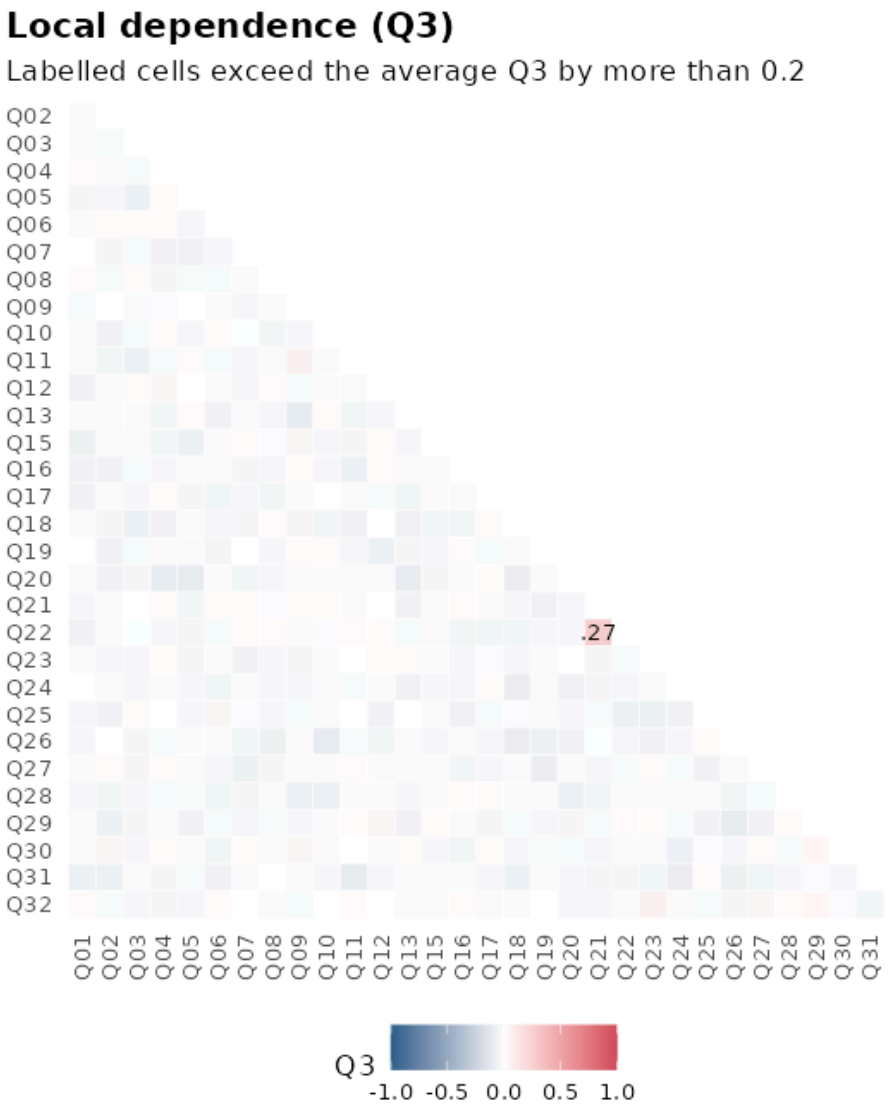
Item fit

Item fit

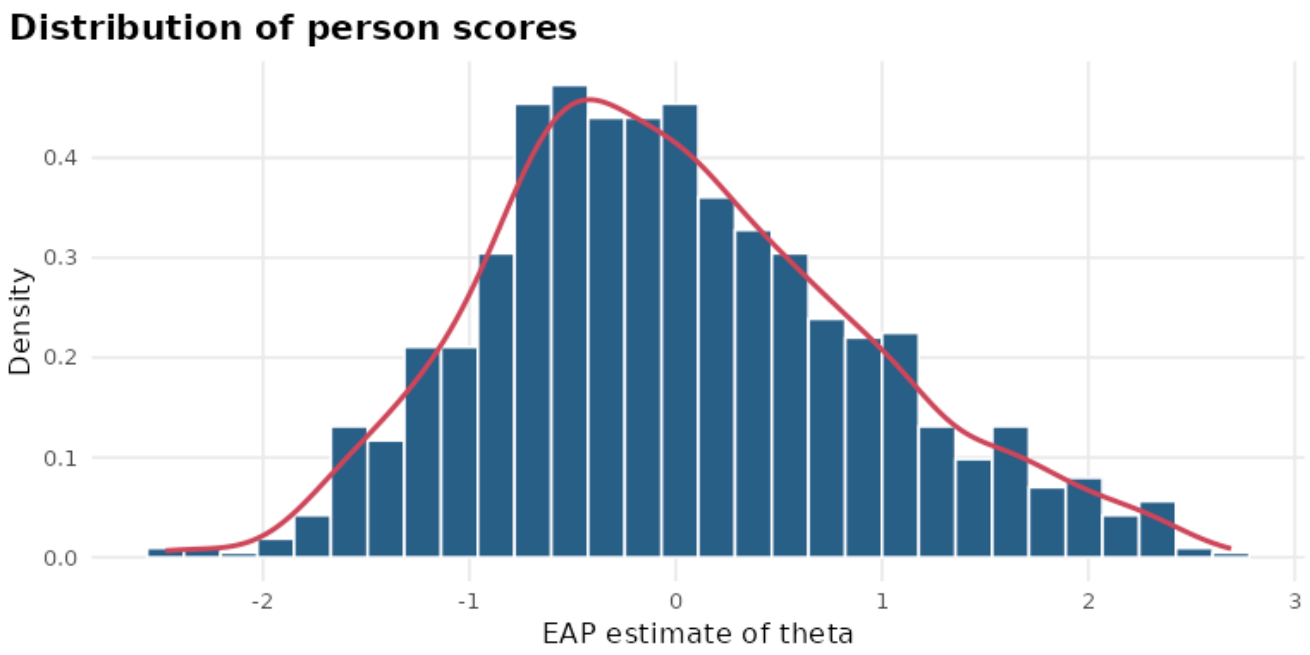
Red points: p < 0.01; dashed line: RMSEA = .05



Local dependence



Scores



Sum score conversion table

Expected a posteriori (EAP) ability for each total score, with its standard error. Observed and expected counts show how well the model reproduces the score distribution.

Sum.Scores	F1	SE_F1	observed	expected
0	-3.073	0.557	0	0.136
1	-2.824	0.532	0	0.670
2	-2.586	0.509	1	1.868
3	-2.362	0.489	1	3.927
4	-2.151	0.470	2	6.956
5	-1.952	0.455	4	10.980
6	-1.763	0.441	11	15.943
7	-1.583	0.430	19	21.725
8	-1.410	0.421	25	28.154
9	-1.243	0.414	41	35.027
10	-1.081	0.408	33	42.125
11	-0.923	0.404	61	49.223
12	-0.767	0.401	70	56.084

Sum.Scores	F1	SE_F1	observed	expected
13	-0.614	0.400	90	62.465
14	-0.462	0.399	94	68.117
15	-0.310	0.400	69	72.792
16	-0.159	0.402	85	76.263
17	-0.006	0.405	82	78.328
18	0.149	0.410	76	78.831
19	0.306	0.416	68	77.660
20	0.466	0.423	59	74.756
21	0.631	0.432	64	70.121
22	0.801	0.442	45	63.826
23	0.976	0.454	40	56.040
24	1.159	0.468	46	47.062
25	1.350	0.484	21	37.350
26	1.550	0.503	30	27.532
27	1.760	0.523	21	18.369
28	1.981	0.545	19	10.654
29	2.212	0.569	17	5.021
30	2.451	0.593	5	1.693
31	2.692	0.615	1	0.301

Item health cards

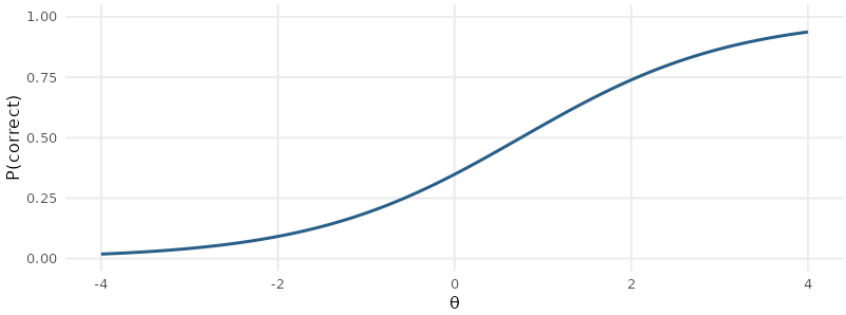
Item health card: Q01

Key statistics

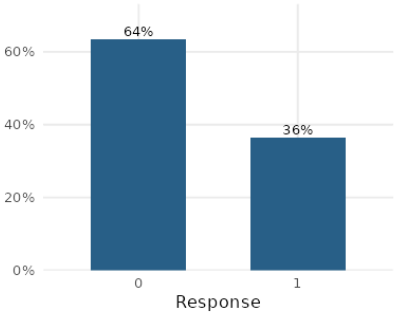
GREEN

Model: 2PL
a = 0.83, b = 0.75
Item-total r = 0.32 Mean = 0.36
S-X2(22) = 14.43, p = .885, RMSEA = 0.00
Flags: none

Trace curves



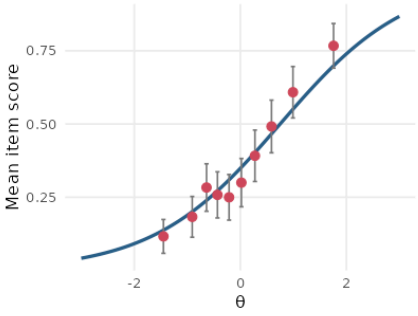
Responses



Information



Observed vs. expected



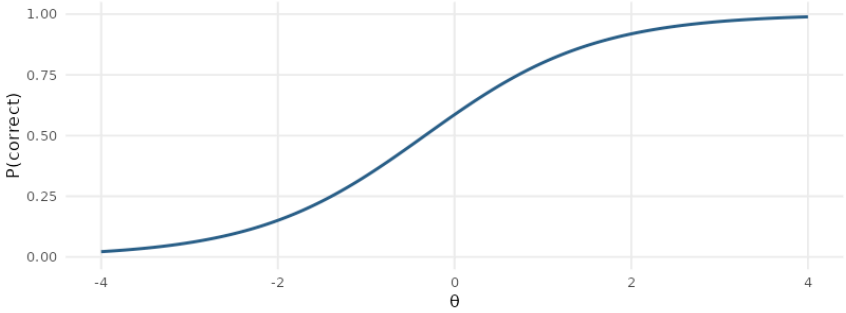
Item health card: Q02

Key statistics

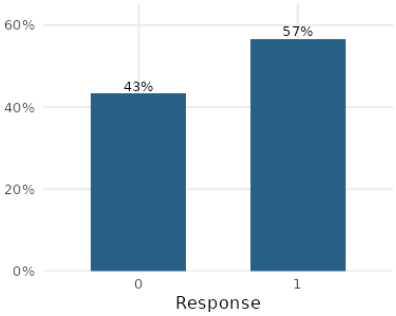
GREEN

Model: 2PL
a = 1.04, b = -0.34
Item-total r = 0.37 Mean = 0.57
S-X2(21) = 25.56, p = .224, RMSEA = 0.01
Flags: none

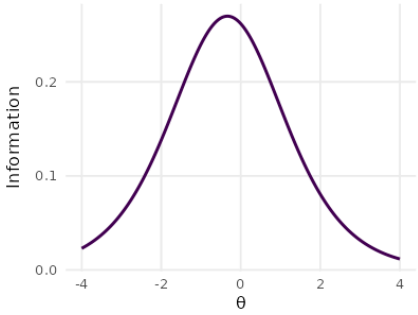
Trace curves



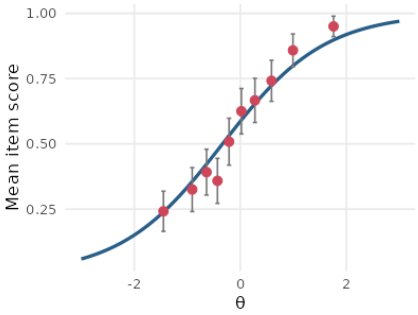
Responses



Information



Observed vs. expected



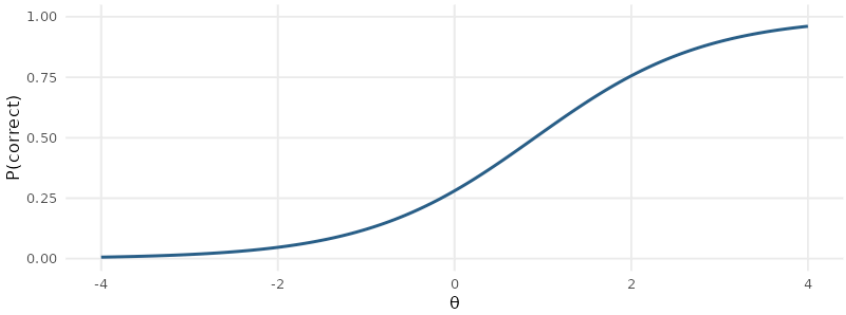
Item health card: Q03

Key statistics

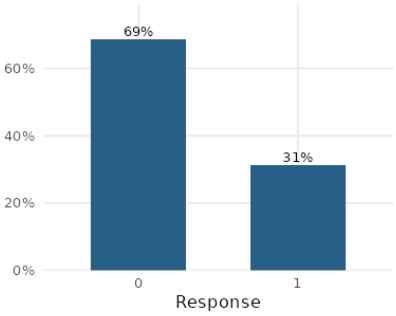
GREEN

Model: 2PL
a = 1.04, b = 0.91
Item-total r = 0.38 Mean = 0.31
S-X2(21) = 22.89, p = .350, RMSEA = 0.00
Flags: none

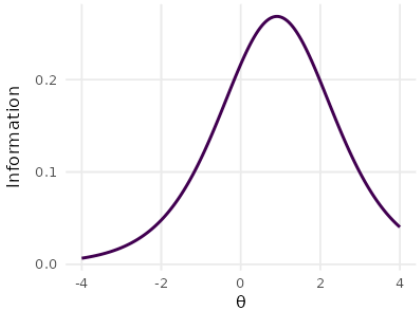
Trace curves



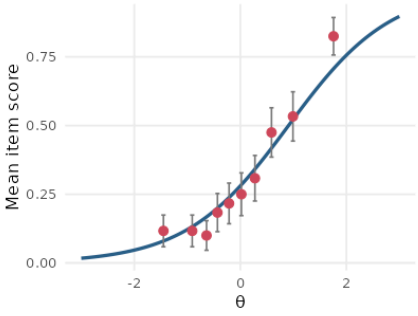
Responses



Information



Observed vs. expected



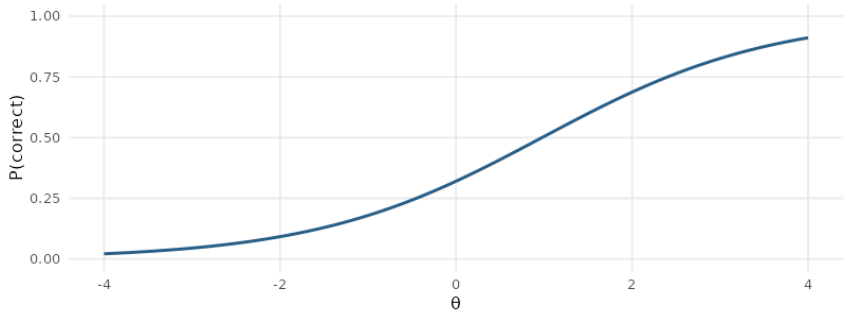
Item health card: Q04

Key statistics

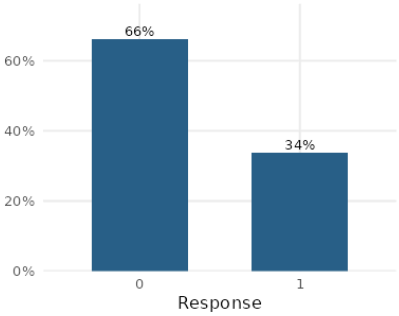
RED

Model: 2PL
a = 0.77, b = 0.98
Item-total r = 0.31 Mean = 0.34
S-X2(22) = 43.36, p = .004, RMSEA = 0.02
Flags: misfit

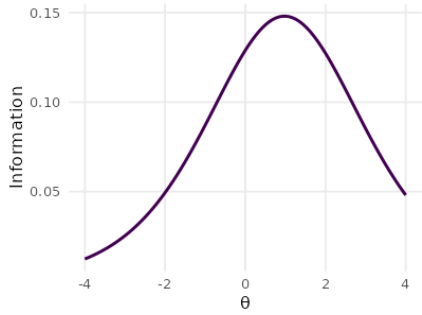
Trace curves



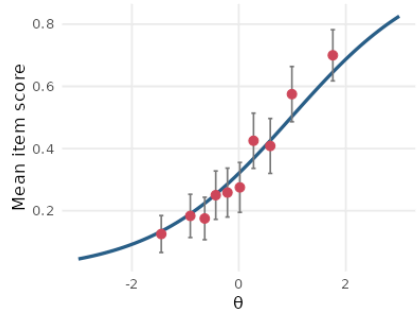
Responses



Information



Observed vs. expected



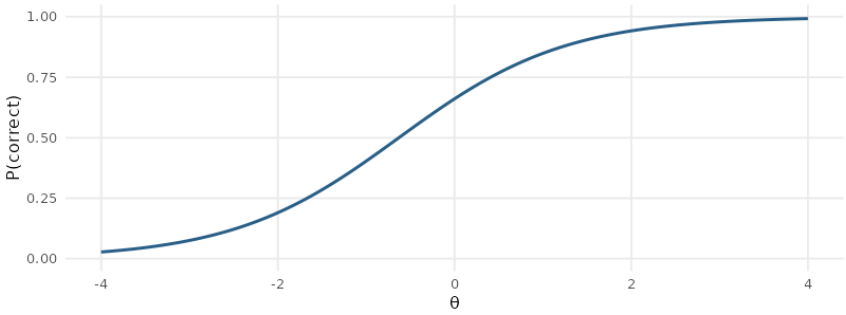
Item health card: Q05

Key statistics

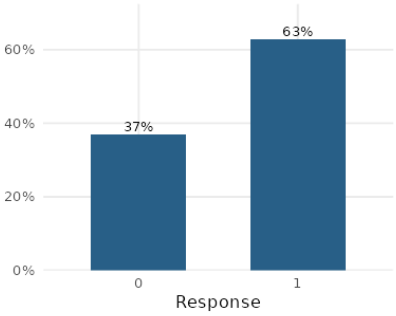
GREEN

Model: 2PL
a = 1.06, b = -0.63
Item-total r = 0.36 Mean = 0.63
S-X2(20) = 20.14, p = .450, RMSEA = 0.00
Flags: none

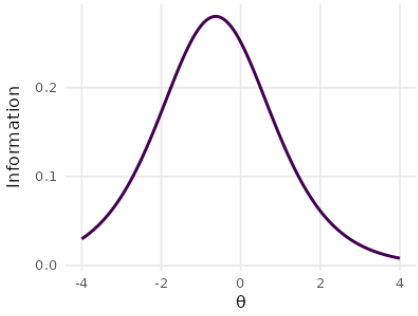
Trace curves



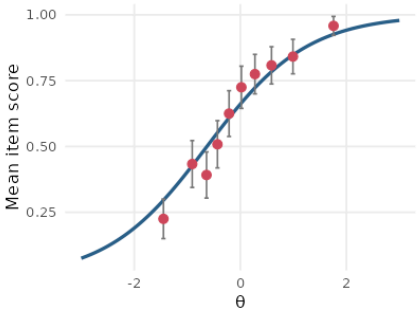
Responses



Information



Observed vs. expected



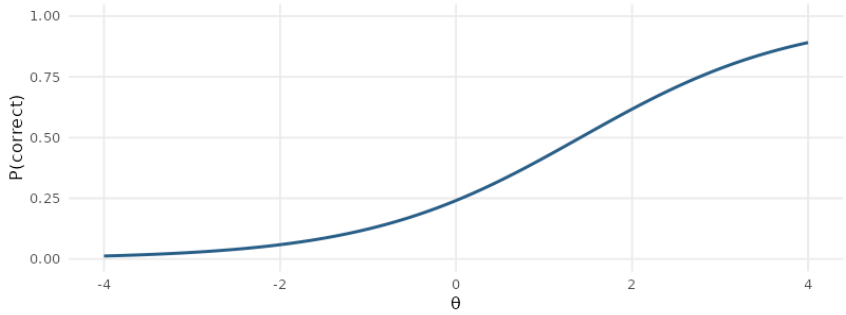
Item health card: Q06

Key statistics

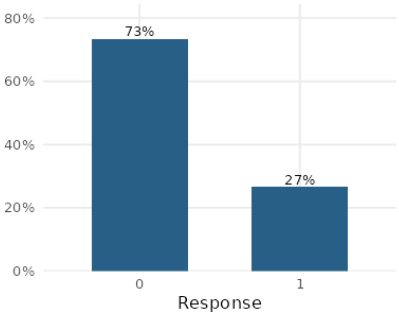
GREEN

Model: 2PL
a = 0.81, b = 1.41
Item-total r = 0.30 Mean = 0.27
S-X2(21) = 34.59, p = .031, RMSEA = 0.02
Flags: none

Trace curves



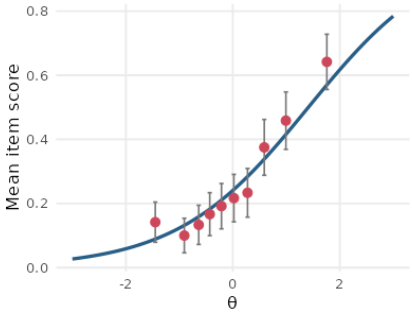
Responses



Information



Observed vs. expected



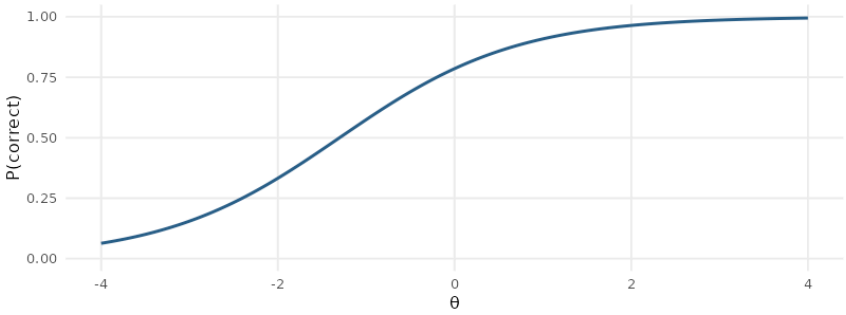
Item health card: Q07

Key statistics

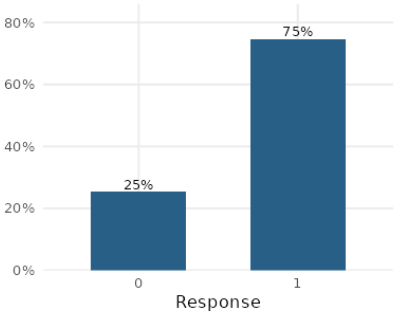
GREEN

Model: 2PL
a = 1.00, b = -1.30
Item-total r = 0.32 Mean = 0.75
S-X2(20) = 23.28, p = .275, RMSEA = 0.01
Flags: none

Trace curves



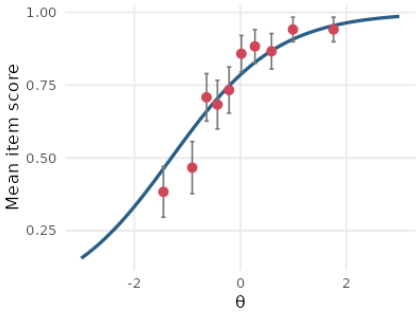
Responses



Information



Observed vs. expected



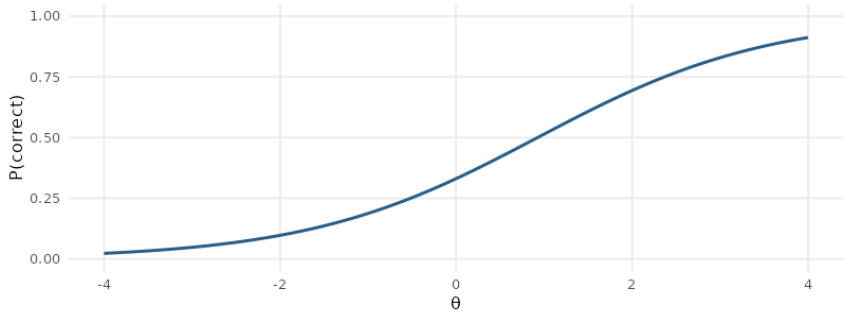
Item health card: Q08

Key statistics

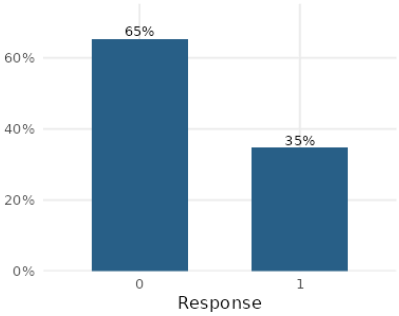
GREEN

Model: 2PL
a = 0.76, b = 0.93
Item-total r = 0.30 Mean = 0.35
S-X2(22) = 16.31, p = .800, RMSEA = 0.00
Flags: none

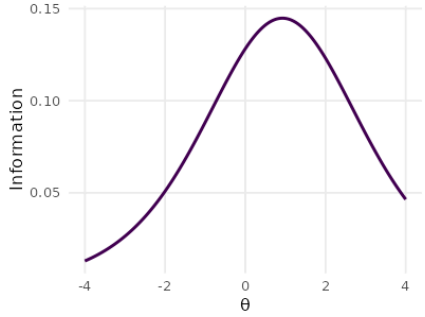
Trace curves



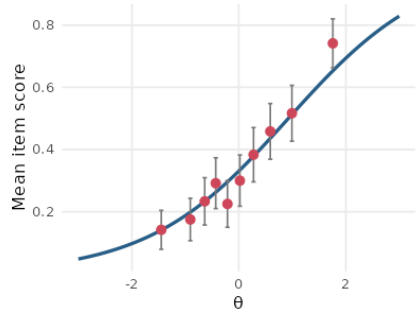
Responses



Information



Observed vs. expected



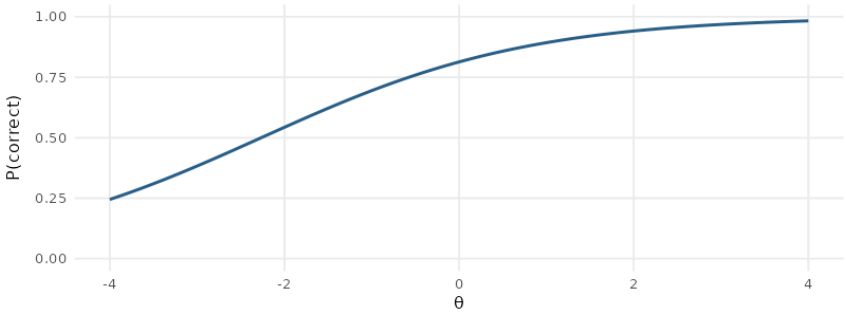
Item health card: Q09

Key statistics

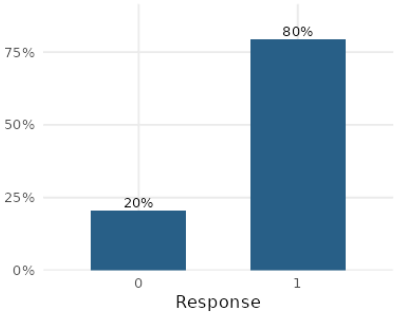
GREEN

Model: 2PL
a = 0.65, b = -2.27
Item-total r = 0.21 Mean = 0.80
S-X2(22) = 28.22, p = .169, RMSEA = 0.01
Flags: none

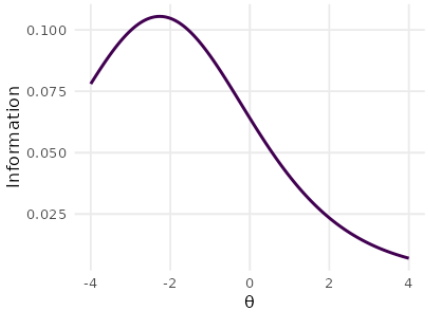
Trace curves



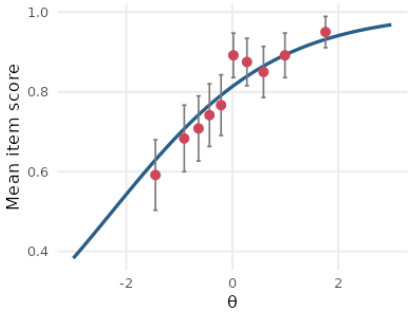
Responses



Information



Observed vs. expected



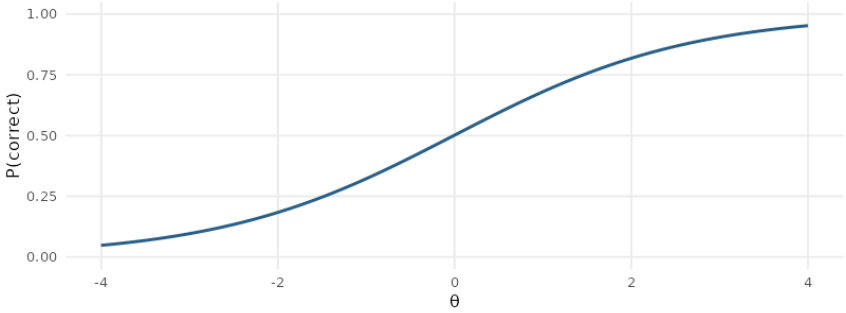
Item health card: Q10

Key statistics

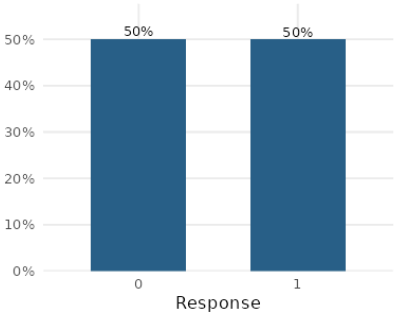
GREEN

Model: 2PL
a = 0.75, b = -0.01
Item-total r = 0.30 Mean = 0.50
S-X2(22) = 31.59, p = .085, RMSEA = 0.01
Flags: none

Trace curves



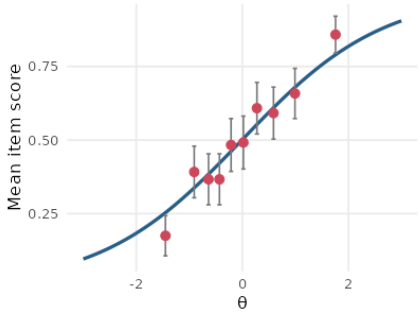
Responses



Information



Observed vs. expected



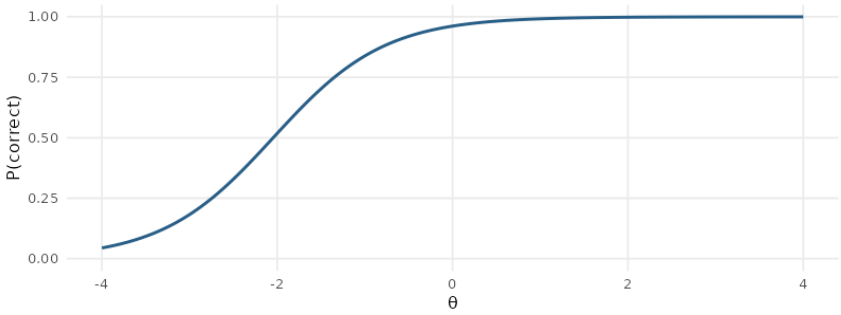
Item health card: Q11

Key statistics

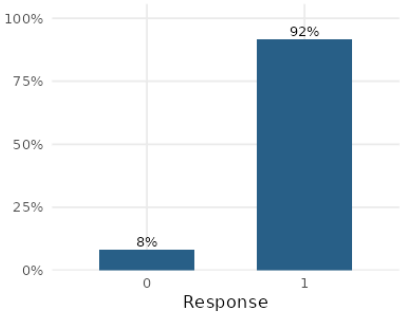
GREEN

Model: 2PL
a = 1.57, b = -2.05
Item-total r = 0.28 Mean = 0.92
S-X2(16) = 21.98, p = .144, RMSEA = 0.01
Flags: none

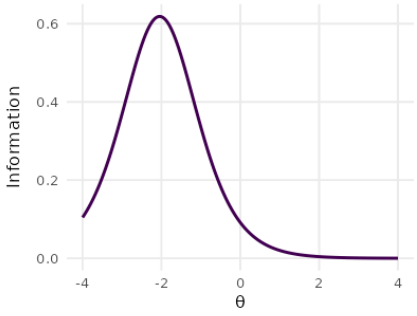
Trace curves



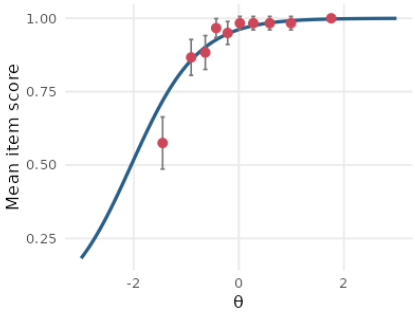
Responses



Information



Observed vs. expected



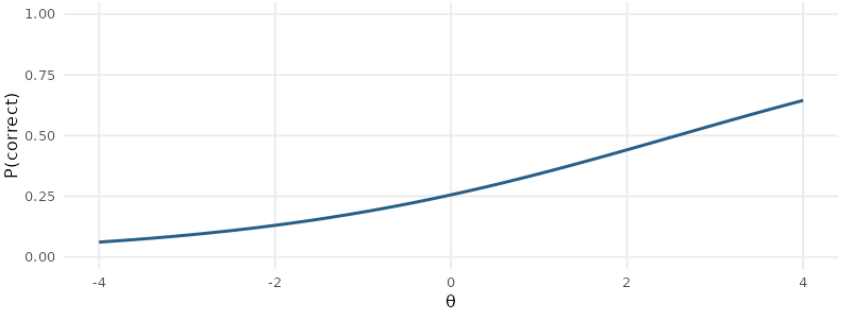
Item health card: Q12

Key statistics

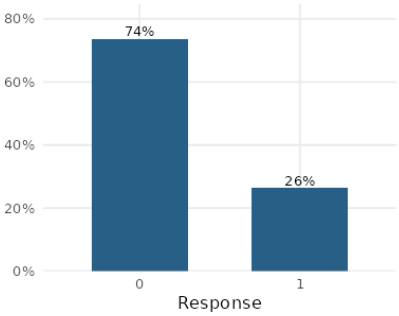
RED

Model: 2PL
a = 0.42, b = 2.57
Item-total r = 0.16 Mean = 0.26
S-X2(22) = 46.58, p = .002, RMSEA = 0.03
Flags: low discrimination; low item-total co

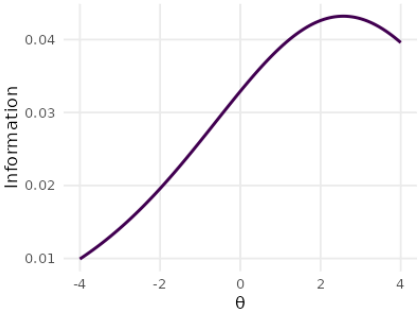
Trace curves



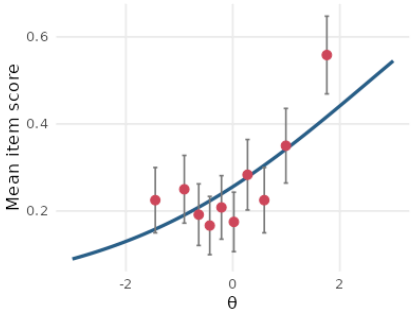
Responses



Information



Observed vs. expected



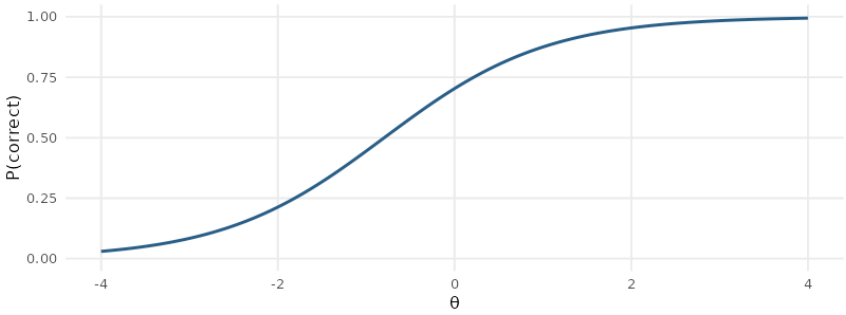
Item health card: Q13

Key statistics

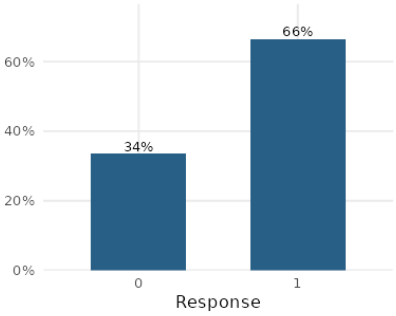
GREEN

Model: 2PL
a = 1.08, b = -0.80
Item-total r = 0.36 Mean = 0.67
S-X2(20) = 30.60, p = .061, RMSEA = 0.02
Flags: none

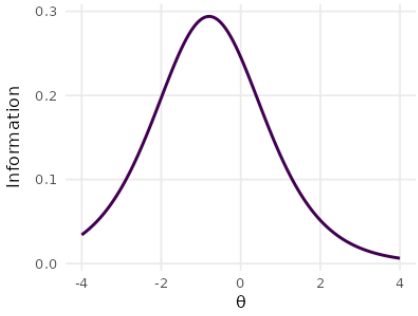
Trace curves



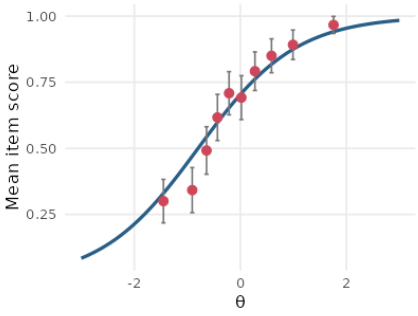
Responses



Information



Observed vs. expected



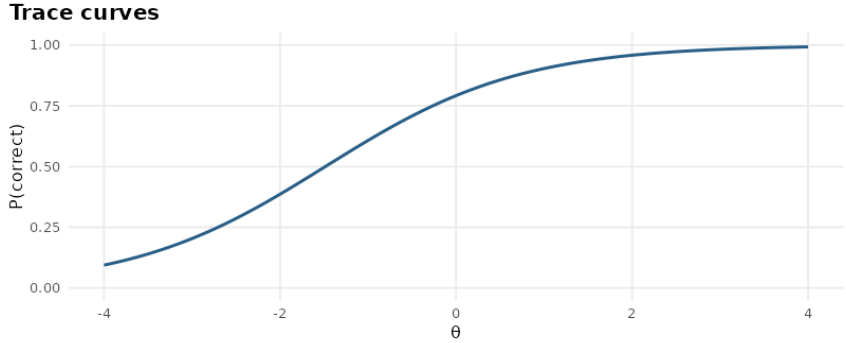
Item health card: Q15

Key statistics

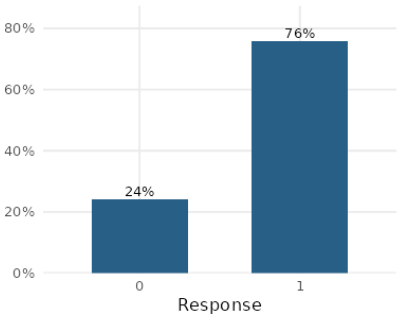
GREEN

Model: 2PL
a = 0.90, b = -1.49
Item-total r = 0.29 Mean = 0.76
S-X2(21) = 26.69, p = .181, RMSEA = 0.01
Flags: none

Trace curves



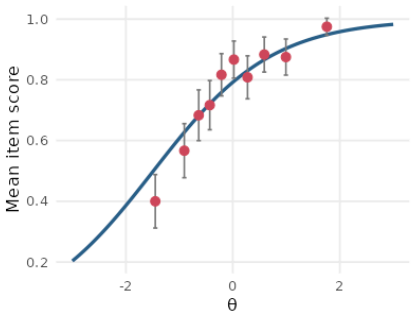
Responses



Information



Observed vs. expected



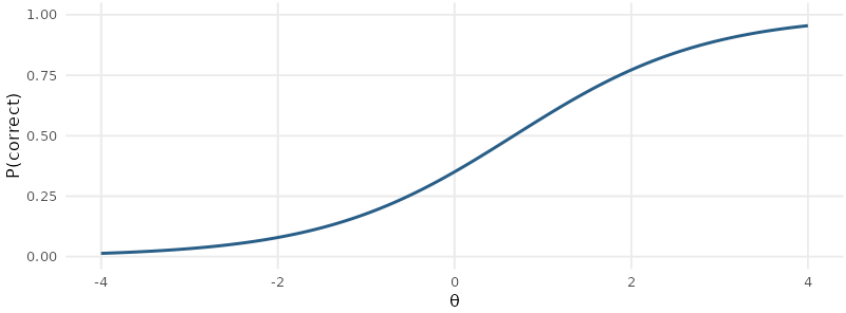
Item health card: Q16

Key statistics

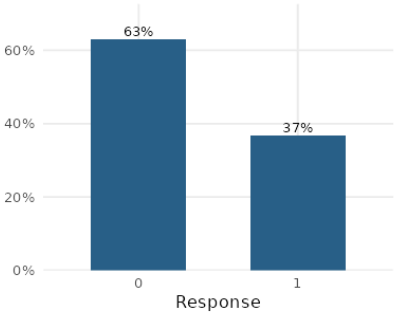
GREEN

Model: 2PL
a = 0.92, b = 0.67
Item-total r = 0.34 Mean = 0.37
S-X2(22) = 30.28, p = .112, RMSEA = 0.01
Flags: none

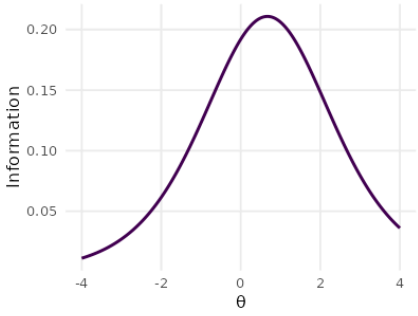
Trace curves



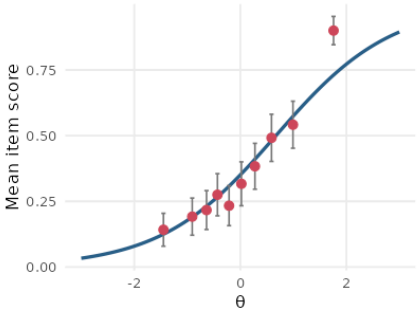
Responses



Information



Observed vs. expected



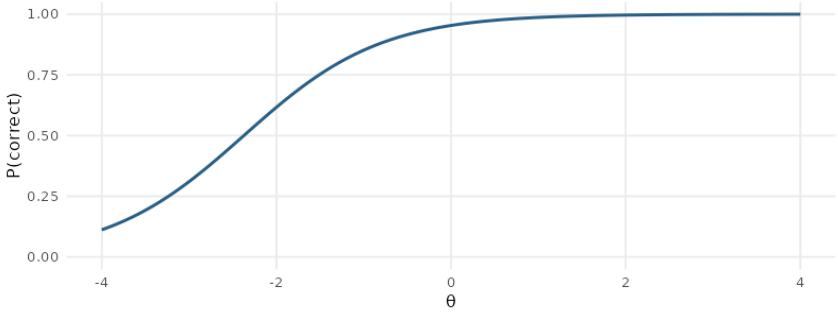
Item health card: Q17

Key statistics

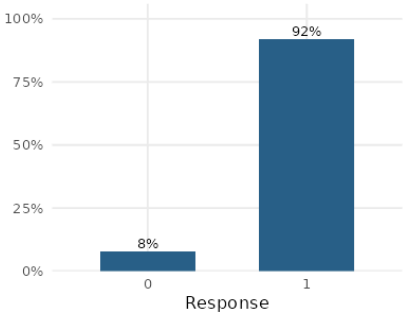
RED

Model: 2PL
a = 1.27, b = -2.37
Item-total r = 0.23 Mean = 0.92
S-X2(17) = 38.05, p = .002, RMSEA = 0.03
Flags: misfit

Trace curves



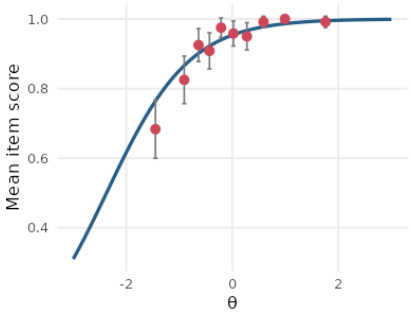
Responses



Information



Observed vs. expected



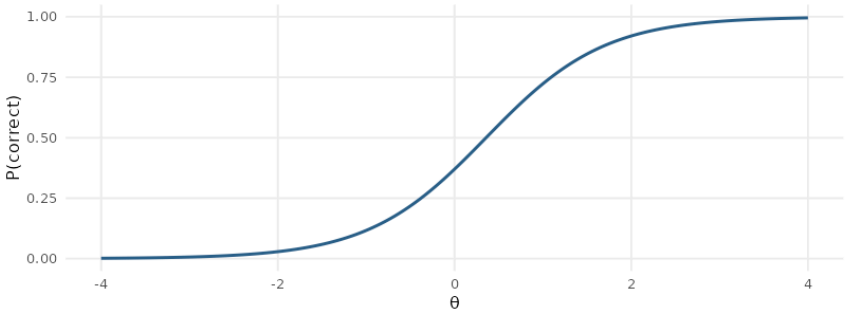
Item health card: Q18

Key statistics

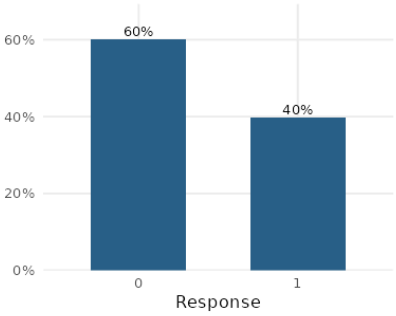
GREEN

Model: 2PL
a = 1.49, b = 0.36
Item-total r = 0.48 Mean = 0.40
S-X2(20) = 32.22, p = .041, RMSEA = 0.02
Flags: none

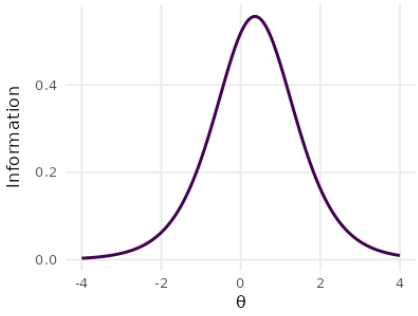
Trace curves



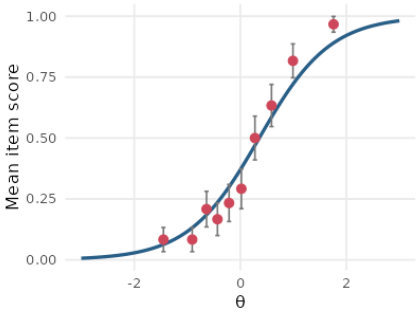
Responses



Information



Observed vs. expected



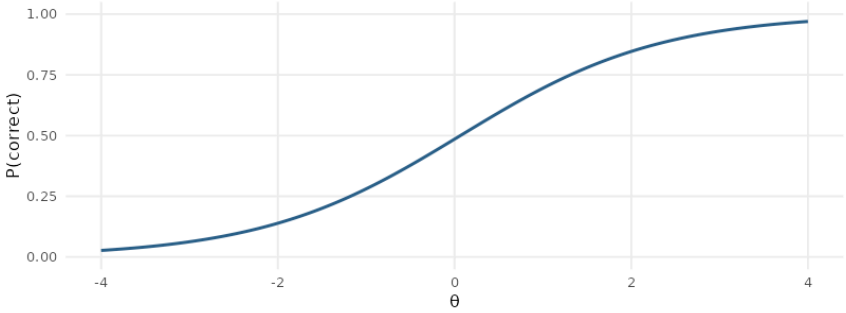
Item health card: Q19

Key statistics

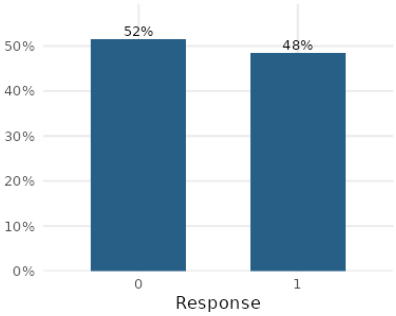
GREEN

Model: 2PL
a = 0.88, b = 0.07
Item-total r = 0.34 Mean = 0.48
S-X2(21) = 15.92, p = .774, RMSEA = 0.00
Flags: none

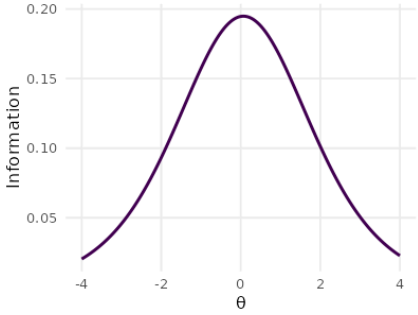
Trace curves



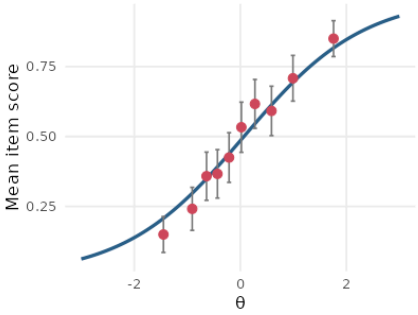
Responses



Information



Observed vs. expected



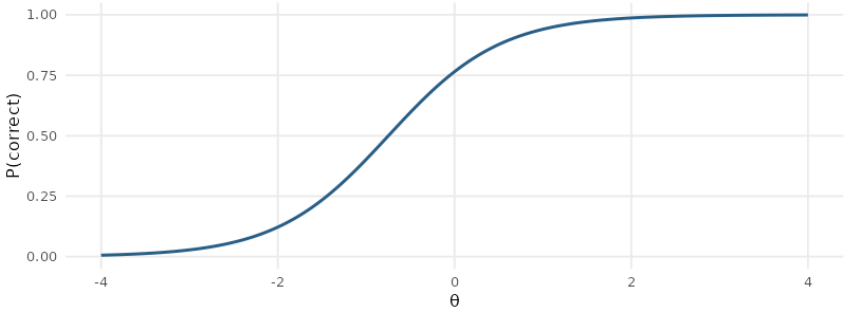
Item health card: Q20

Key statistics

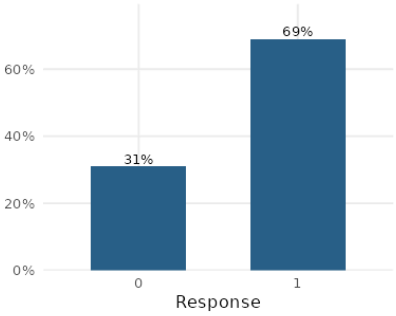
GREEN

Model: 2PL
a = 1.58, b = -0.75
Item-total r = 0.44 Mean = 0.69
S-X2(18) = 24.10, p = .152, RMSEA = 0.01
Flags: none

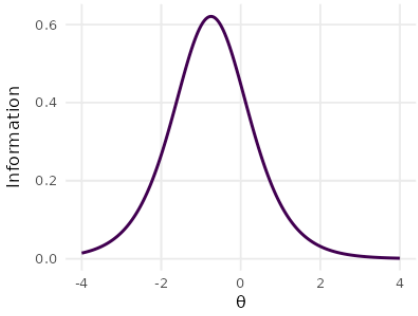
Trace curves



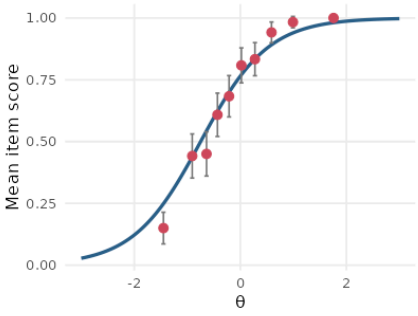
Responses



Information



Observed vs. expected



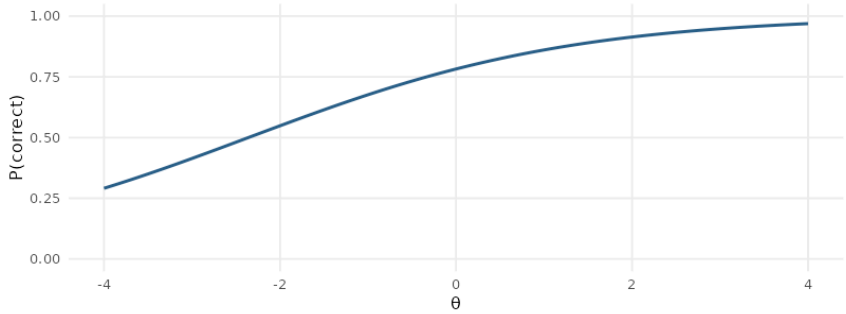
Item health card: Q21

Key statistics

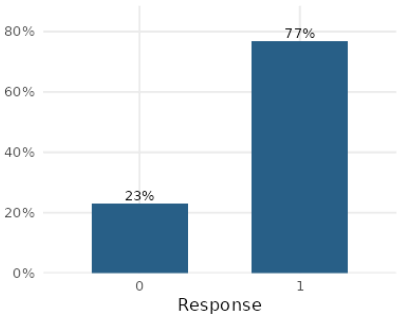
AMBER

Model: 2PL
a = 0.54, b = -2.36
Item-total r = 0.19 Mean = 0.77
S-X2(23) = 23.69, p = .421, RMSEA = 0.00
Flags: low item-total correlation

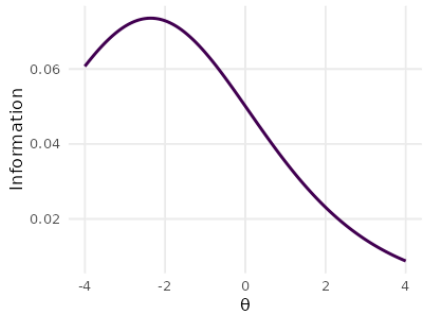
Trace curves



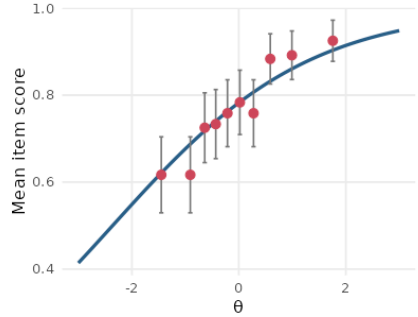
Responses



Information



Observed vs. expected



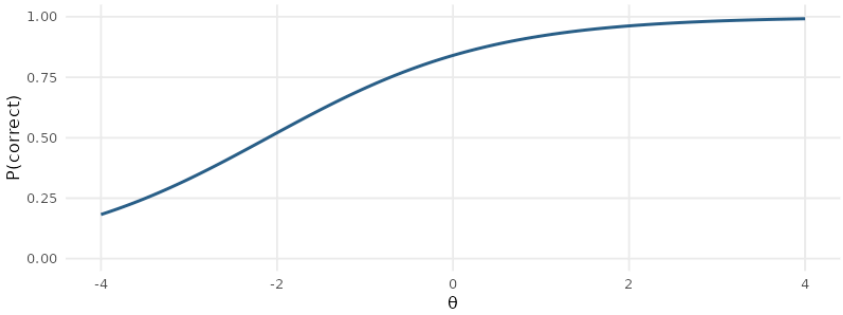
Item health card: Q22

Key statistics

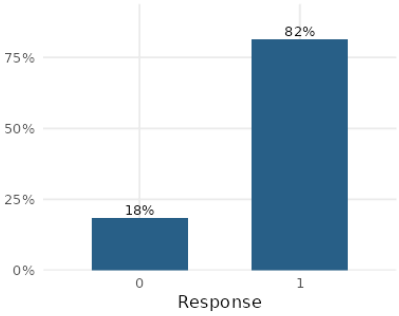
GREEN

Model: 2PL
a = 0.79, b = -2.10
Item-total r = 0.24 Mean = 0.81
S-X2(21) = 16.52, p = .740, RMSEA = 0.00
Flags: none

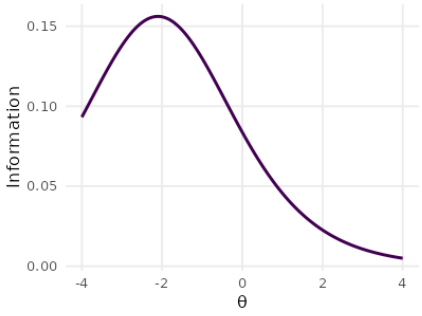
Trace curves



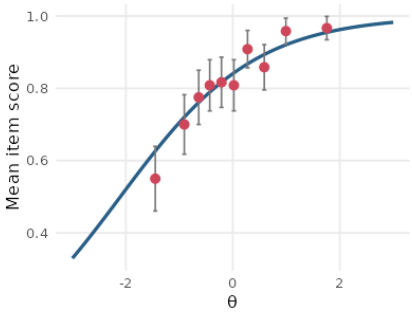
Responses



Information



Observed vs. expected



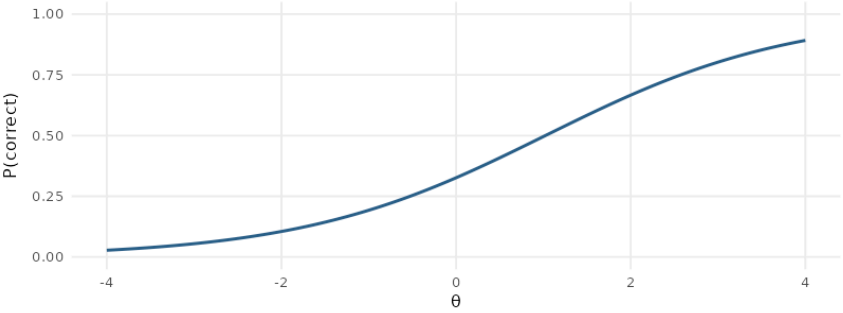
Item health card: Q23

Key statistics

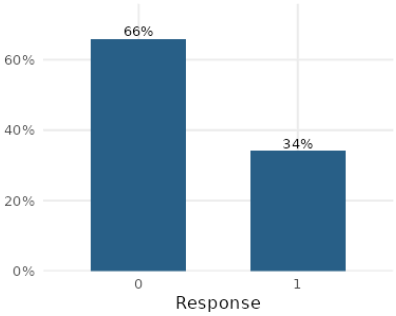
GREEN

Model: 2PL
a = 0.71, b = 1.02
Item-total r = 0.28 Mean = 0.34
S-X2(22) = 16.36, p = .798, RMSEA = 0.00
Flags: none

Trace curves



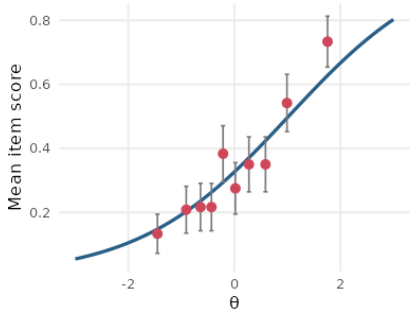
Responses



Information



Observed vs. expected



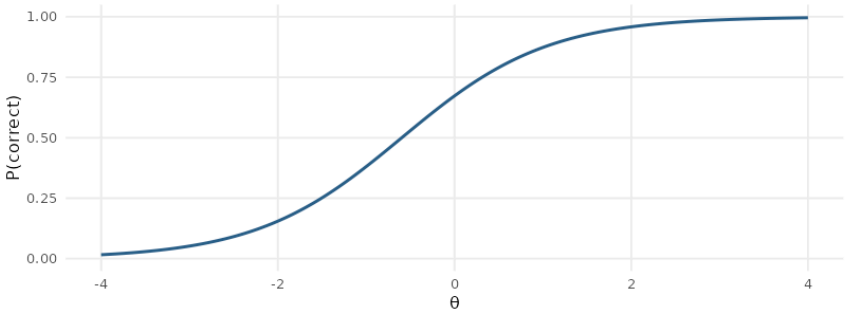
Item health card: Q24

Key statistics

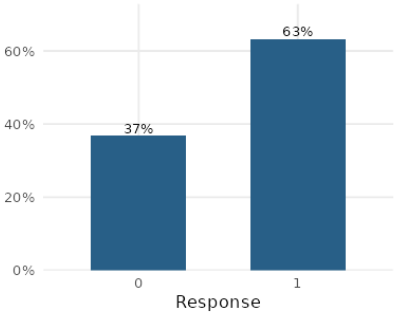
GREEN

Model: 2PL
a = 1.21, b = -0.60
Item-total r = 0.39 Mean = 0.63
S-X2(20) = 17.98, p = .588, RMSEA = 0.00
Flags: none

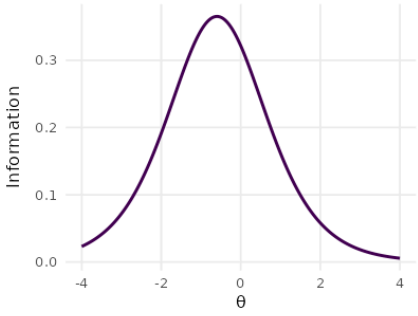
Trace curves



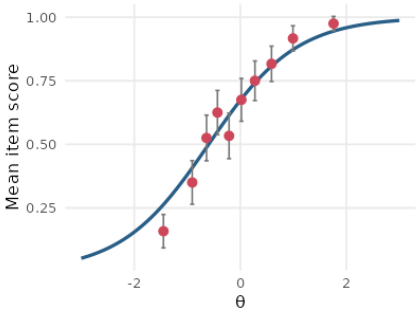
Responses



Information



Observed vs. expected



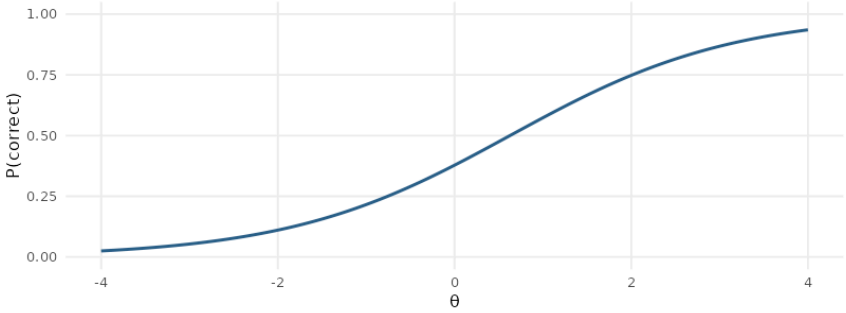
Item health card: Q25

Key statistics

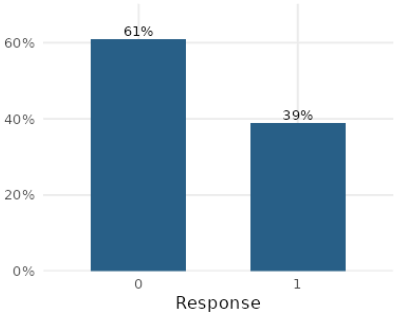
GREEN

Model: 2PL
a = 0.79, b = 0.63
Item-total r = 0.30 Mean = 0.39
S-X2(22) = 22.18, p = .449, RMSEA = 0.00
Flags: none

Trace curves



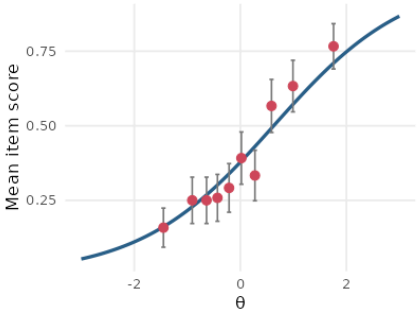
Responses



Information



Observed vs. expected



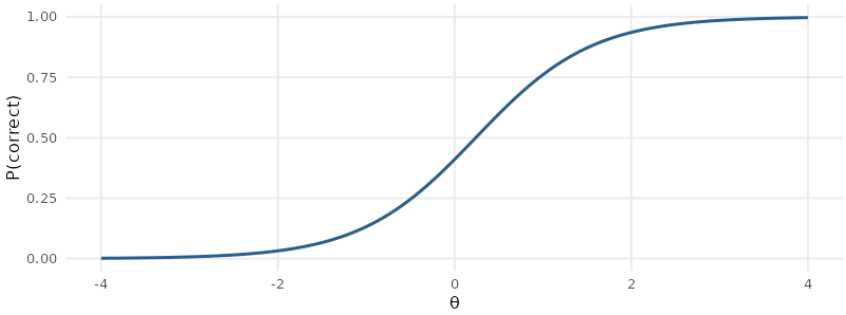
Item health card: Q26

Key statistics

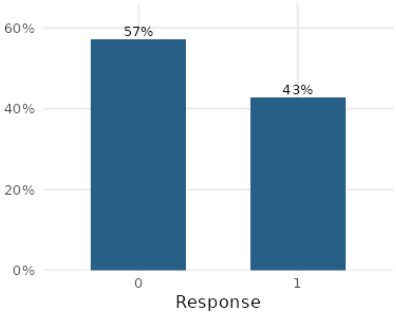
GREEN

Model: 2PL
a = 1.52, b = 0.24
Item-total r = 0.48 Mean = 0.43
S-X2(19) = 24.10, p = .193, RMSEA = 0.01
Flags: none

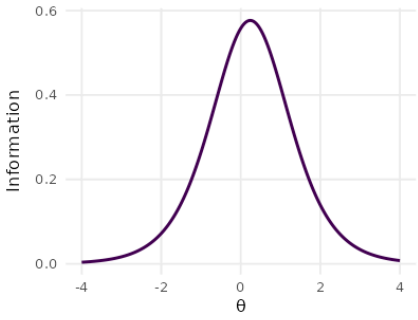
Trace curves



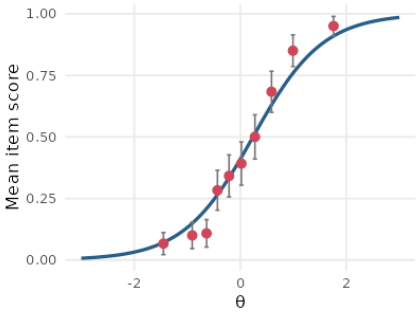
Responses



Information



Observed vs. expected



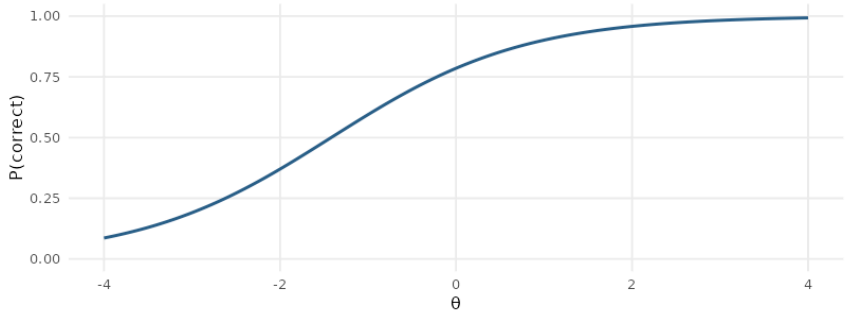
Item health card: Q27

Key statistics

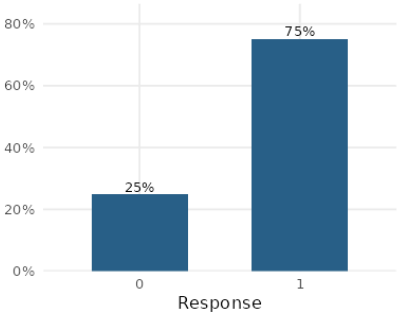
GREEN

Model: 2PL
a = 0.91, b = -1.42
Item-total r = 0.29 Mean = 0.75
S-X2(21) = 16.81, p = .722, RMSEA = 0.00
Flags: none

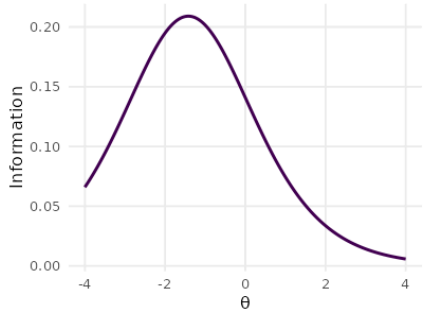
Trace curves



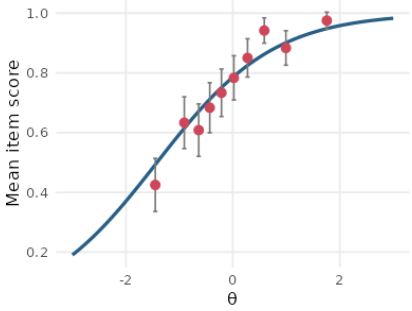
Responses



Information



Observed vs. expected



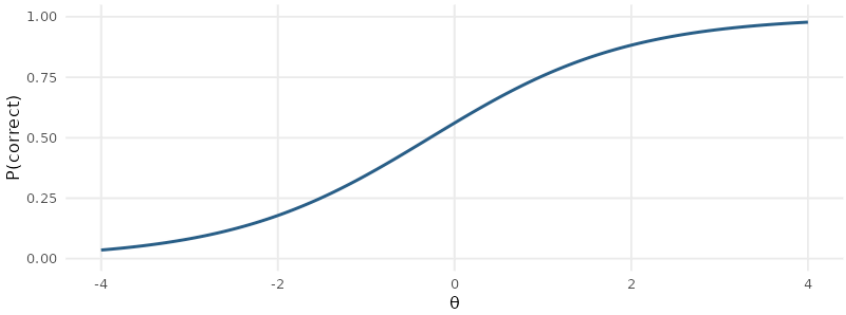
Item health card: Q28

Key statistics

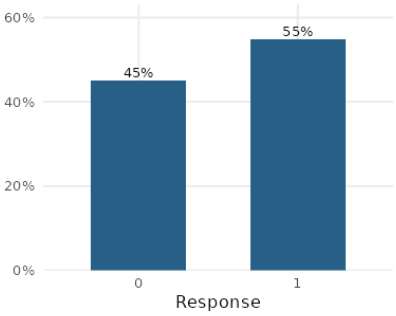
GREEN

Model: 2PL
a = 0.89, b = -0.28
Item-total r = 0.33 Mean = 0.55
S-X2(21) = 17.98, p = .650, RMSEA = 0.00
Flags: none

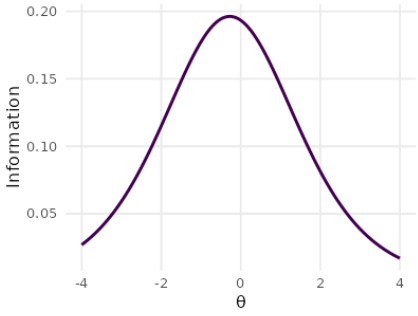
Trace curves



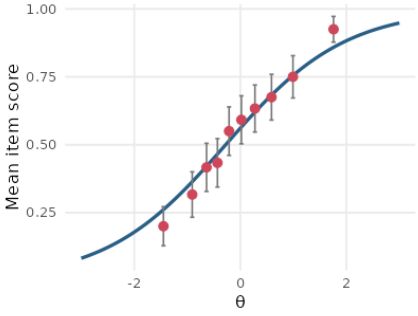
Responses



Information



Observed vs. expected



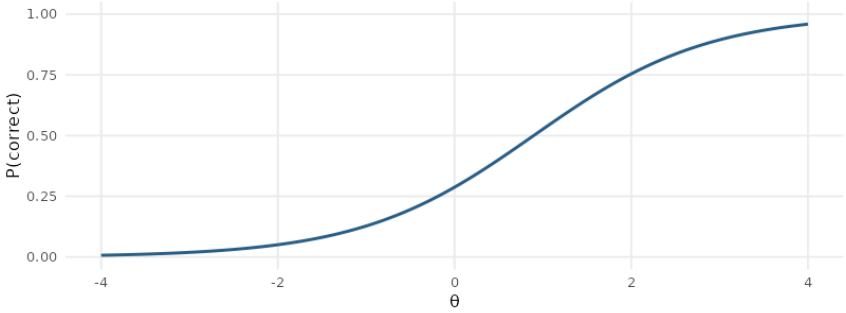
Item health card: Q29

Key statistics

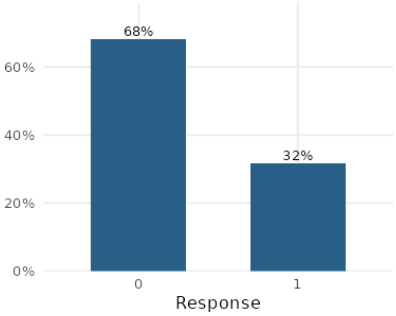
GREEN

Model: 2PL
a = 1.01, b = 0.90
Item-total r = 0.37 Mean = 0.32
S-X2(21) = 26.80, p = .178, RMSEA = 0.01
Flags: none

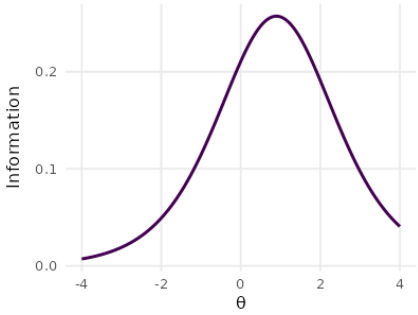
Trace curves



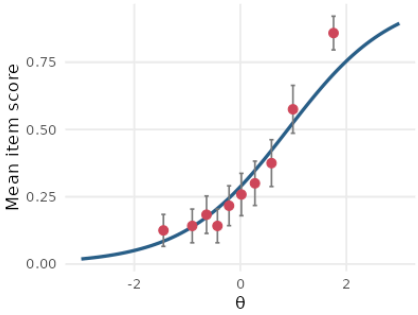
Responses



Information



Observed vs. expected



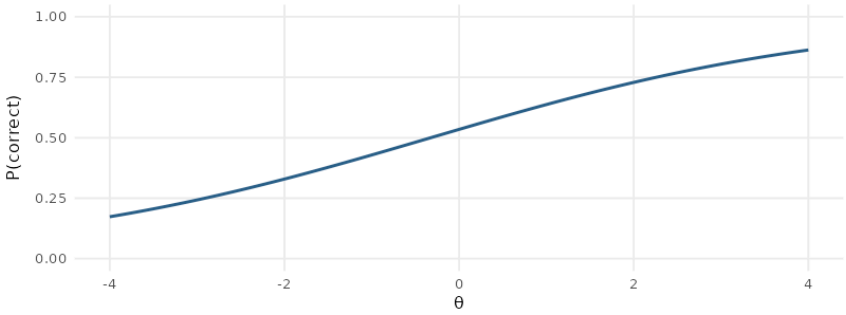
Item health card: Q30

Key statistics

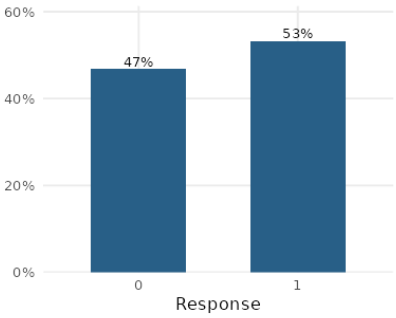
RED

Model: 2PL
a = 0.43, b = -0.32
Item-total r = 0.19 Mean = 0.53
S-X2(23) = 35.34, p = .048, RMSEA = 0.02
Flags: low discrimination; low item-total co

Trace curves



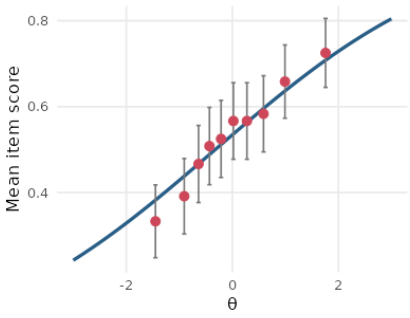
Responses



Information



Observed vs. expected



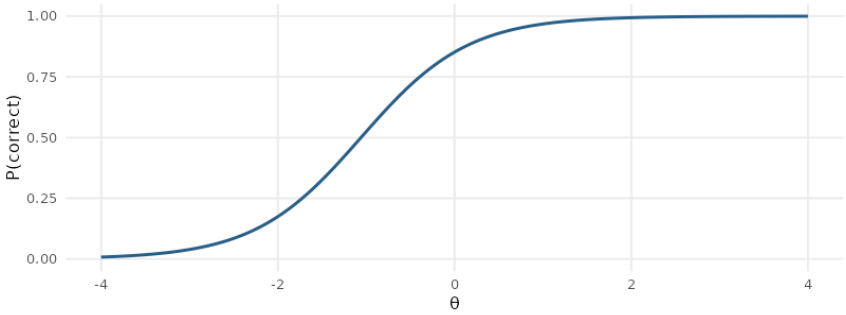
Item health card: Q31

Key statistics

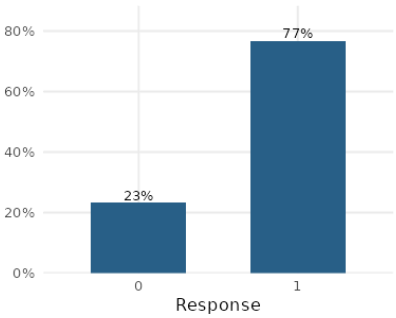
GREEN

Model: 2PL
a = 1.65, b = -1.06
Item-total r = 0.41 Mean = 0.77
S-X2(17) = 11.50, p = .829, RMSEA = 0.00
Flags: none

Trace curves



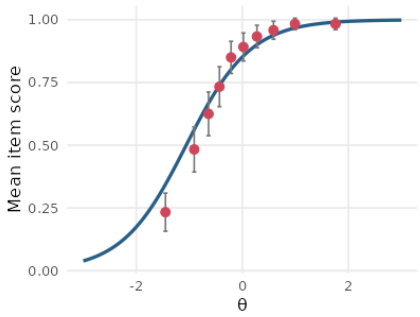
Responses

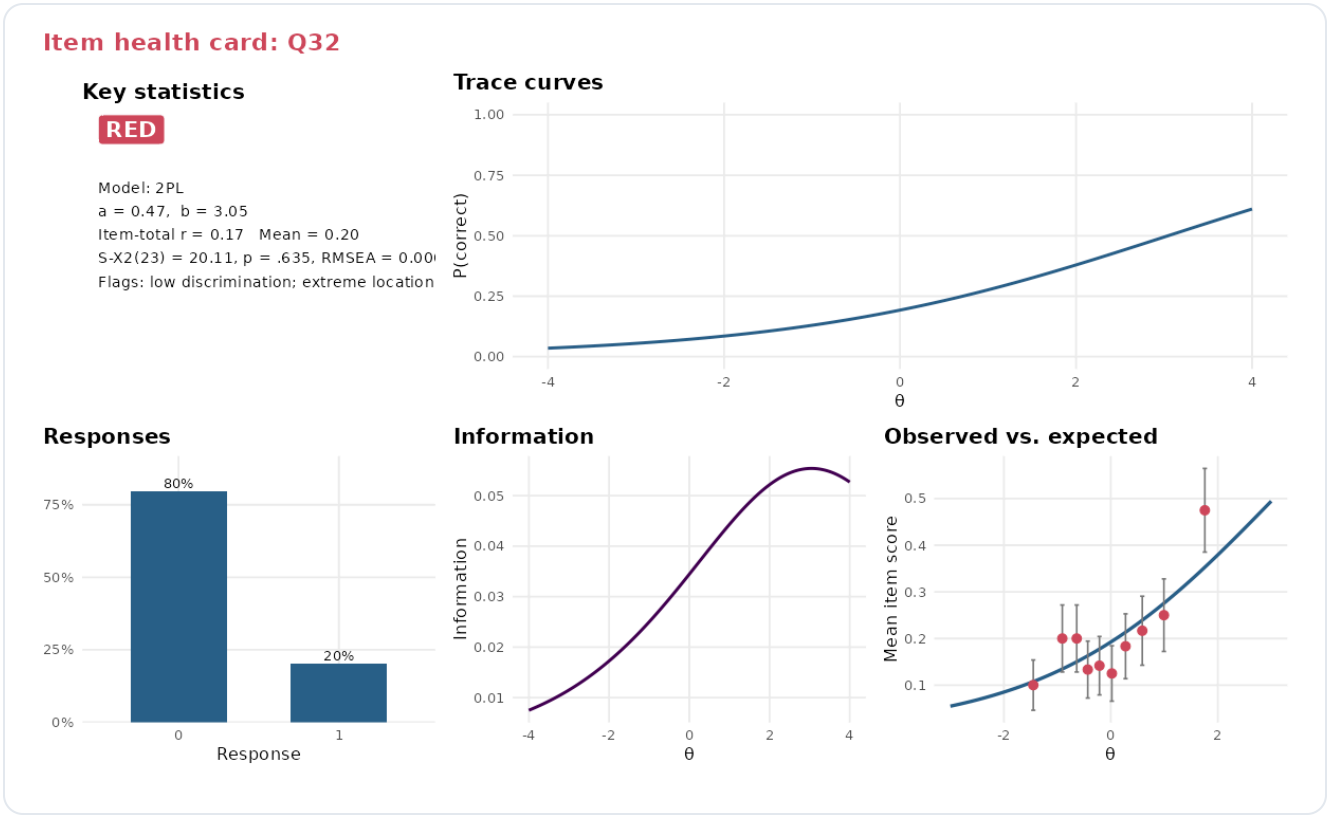


Information



Observed vs. expected





Blueprint

Construct: Professional knowledge for certification. Test takers: not stated. Score use: pass/fail certification. Forms have 12 items (multiple choice); target reliability 0.85.

content	cognitive	n
Ethics	NA	4
Practice	NA	5
Law	NA	3

Field-test design

The field test uses 2 form(s) of 18 items, each taken by 600 test takers, with 4 anchor items on every form, for calibration with the 2PL model.

In 2 simulated field tests the slopes were recovered with an RMSE of 0.22 and the intercepts with an RMSE of 0.12.

form	role	item
1	anchor	4
2	anchor	4
1	unique	14
2	unique	14

Groups

Ability estimates (EAP) by group, with the average standard error. Groups with fewer than 10 respondents are left out.

group	n	mean_theta	sd_theta	mean_se
English	750	0.118	0.913	0.411
French	450	-0.198	0.878	0.402

Cut scores

1 cut score(s) on the summed-score scale divide test takers into Fail, Pass. Decision quality was estimated by the method of Lee (2010) for the calibrated test, over 1,200 abilities: classification accuracy 0.861, consistency 0.809 (kappa 0.61).

Cuts

cut	label	theta	raw	sem
1	Pass	0.157	18	0.421

Decision quality by cut

cut	label	accuracy	consistency	kappa	false_positive	false_negative
1	Pass	0.861	0.809	0.613	0.095	0.044

Expected share in each category

category	expected	true
Fail	0.555	0.606
Pass	0.445	0.394

Scoring

Test takers were scored by EAP: mean ability -0.00 (SD 0.91), median standard error 0.39.

Test security

Security checks run: preknowledge, answer copying. Security flags are evidence for an investigation, not proof of misconduct; individual test takers are not named in this report.

No test taker showed evidence.

Pipeline decisions

Every stage ran its checks and raised flags for review. Flags only warned (strict = FALSE). 9 flag(s) were raised; 1 are still open and 8 were acknowledged, overridden or fixed.

Stages

stage	status	summary
0 Blueprint	done	3 cells, 12 items per form
1 Item generation	not run	
2 Item review	not run	
3 Field-test design	done	2 forms of 18 items, 4 anchors, 1,200 test takers
4 Data intake	done	1,200 respondents, 32 items, scored with key
5 Calibration	done	2PL, converged, reliability 0.83
6 Linking and bank	done	31 items added to the bank
7 Forms	not run	

stage	status	summary
8 Cut scores	done	1 cut, accuracy 0.86, consistency 0.81
9 Scoring	done	EAP scores for 1,200 test takers, reliability 0.83
10 Test security	done	2 checks; 0 test takers with evidence
11 Reporting	done	written to certification-technical-report.html
12 Bank monitoring	not run	

Flags

id	level	stage	check	target	value
2	BLOCK	intake	possible_miskey	Q14	p = 0.18, item-total r = -0.23
4	BLOCK	calibration	dif_large	Q09	MH D-DIF = -2.01, French vs E
1	REVIEW	field_test	recovery	design	RMSE of slopes = 0.22
3	REVIEW	calibration	local_dependence	Q21 & Q22	Q3 = 0.27
5	REVIEW	calibration	dif_moderate	Q27	MH D-DIF = -1.01, French vs E
6	REVIEW	calibration	low_reliability	test	marginal reliability = 0.83
8	REVIEW	cuts	cut_precision	cut 1 (Pass)	SEM = 0.42 at theta = 0.16
9	REVIEW	scoring	score_reliability	scores	empirical reliability = 0.83
7	INFO	calibration	person_misfit	22 respondents	1.8% of respondents

Decision log

time	action	detail
2026-11-03 05:21:32	set blueprint	3 cells; use: pass/fail certification
2026-11-03 05:21:34	design field test	2 forms of 18 items, 4 anchors, 1,200 test takers
2026-11-03 05:21:34	acknowledge	acknowledged #1 recovery (design): Sample fixed by the te
2026-11-03 05:21:34	add responses	1200 respondents, 32 items
2026-11-03 05:21:34	drop items	Q14: Key confirmed wrong by the item writer
2026-11-03 05:21:37	calibrate	2PL
2026-11-03 05:21:37	acknowledge	acknowledged #3 local_dependence (Q21 & Q22): Q21 and

time	action	detail
2026-11-03 05:21:37	override	overridden #4 dif_large (Q09): Translation reviewed; Frenc
2026-11-03 05:21:37	acknowledge	acknowledged #5 dif_moderate (Q27): Sent to the bias pan
2026-11-03 05:21:37	acknowledge	acknowledged #6 low_reliability (test): Operational forms w
2026-11-03 05:21:37	bank items	31 items added to Certification bank
2026-11-03 05:21:37	set cuts	raw cuts at 18
2026-11-03 05:21:37	score persons	EAP
2026-11-03 05:21:37	acknowledge	acknowledged #8 cut_precision (cut 1 (Pass)): Add items ne
2026-11-03 05:21:37	acknowledge	acknowledged #9 score_reliability (scores): Field-test score
2026-11-03 05:21:39	check security	preknowledge, answer_copying
2026-11-03 05:21:40	report	pkgdown/assets/case-study/certification-technical-report.h

Respondents flagged

Counts only; the respondents are listed in p$persons in R.

check	respondents
person_misfit	22

Rules in effect

Preset: standards; 1 rule(s) changed by the user.

▶ Show all rules

Reproducibility

Run this code to reproduce the analysis:

```
library(mirtverse)
p <- mv_pipeline(title = "Certification exam: Form A field test")
p <- set_blueprint(p, cells = c(Ethics = 4, Practice = 5, Law = 3), construct = "Pr
    use = "pass/fail certification", reliability = 0.85)
p <- design_field_test(p, n_forms = 2, n_per_form = 600, items = 32, reps = 2, seed
p <- acknowledge(p, check = "recovery", note = "Sample fixed by the testing window"
```

```
p <- add_responses(p, data = certexam, items = 3:34, key = certexam_key, group = "l
    id = "id")
p <- drop_items(p, items = "Q14", note = "Key confirmed wrong by the item writer")
p <- calibrate(p, model = "2PL")
p <- acknowledge(p, check = "local_dependence", note = "Q21 and Q22 share a case; k
p <- override(p, check = "dif_large", note = "Translation reviewed; French version
p <- acknowledge(p, check = "dif_moderate", note = "Sent to the bias panel")
p <- acknowledge(p, check = "low_reliability", note = "Operational forms will be lo
p <- bank_items(p, name = "Certification bank")
p <- set_cuts(p, cuts = 18, scale = "raw", labels = c("Fail", "Pass"))
p <- score_persons(p)
p <- acknowledge(p, check = "cut_precision", note = "Add items near the cut to the
p <- acknowledge(p, check = "score_reliability", note = "Field-test scores are not
p <- check_security(p, leaked = c("Q01", "Q02", "Q03", "Q04", "Q05"))
p <- pipeline_report(p, file = "pkgdown/assets/case-study/certification-technical-r
    title = "Certification exam, Form A: Technical report", author = "Psychometrics
    open = FALSE)
```

▶ Session information