Item_Response_Theory

March 12, 2024

1 Detailed Question Analysis: *Item Response Theory (IRT)*

- IRT models the complex interplay between individual test takers and question items
- Probability of a correct resopnse is a function of a latent "DQL ability" variable (Θ) and each question's *discrimination* and *difficulty*
- Some items are more informative than others, and at different levels of test taker latent ability
- Similarities with factor analysis & logit/probit models

1.0.1 Load libraries, read data

[152]:
```
options(warn = -1)  #warnings clutter presentation, normally not a good practice
library(ggplot2)
options(repr.plot.width=12, repr.plot.height=8)

library(mirt)  # one of several excellent libraries in R for doing IRT
library(ggmirt) # extension to mirt that allows ggplot2 visualization of models
#Read csv file with question cores as correct/wrong (1/0) integers
qscores <- read.csv("C:/Users/brian_local/PS_demo/qscores.csv", colClasses =␣
 ↪c("integer"))
```

1.0.2 Fit the IRT Model: 2PL, difficulty (a) and discrimination (b), but no guessing, because guessing is presumably impossible here

[153]:
```
unimodel <- 'F1 = 1-13'
fit2PL <- mirt(data = qscores,
               model = unimodel,  # Explain what that means here eventually
               itemtype = "2PL",
               verbose = FALSE)
```

1.0.3 Factor loadings (F1) meausre strength of relationship between item factors and posited latent factor; h2 is F1^2 and represents variance in each item accounted for by latent factor

Relationships are all strong, which is good in general, but implies high correlation among items, which implies redundant questions and makes model estimation more difficult

[154]:
```
summary(fit2PL)
```

```
       F1    h2
Q2  0.978 0.956
Q3  0.990 0.981
Q4  0.975 0.951
Q5  0.964 0.930
Q6  0.972 0.944
Q7  0.967 0.936
Q8  0.961 0.924
Q9  0.988 0.977
Q10 0.997 0.994
Q11 0.966 0.933
Q12 0.979 0.959
Q13 0.901 0.812
Q14 0.946 0.896

SS loadings:  12.192
Proportion Var:  0.938

Factor correlations:

   F1
F1  1
```

1.1 Main Event: IRT parameters

Presents differentiation levels (a), where steep slope = better differentiation, and difficulty (b), which shows theta level that corresponds with a .50 probability of correct response. (Guessing – g – is pre-specified to be zero because if takers need to write code and grading is all-or-none, then guessing is essentially impossible.)

[155]:
```
params2PL <- coef(fit2PL, IRTpars = TRUE, simplify = TRUE)
round(params2PL$items, 2)
```

A matrix: 13 × 4 of type dbl

	a	b	g	u
Q2	7.98	-1.55	0	1
Q3	12.14	-1.49	0	1
Q4	7.50	-1.53	0	1
Q5	6.21	-1.44	0	1
Q6	6.98	-1.14	0	1
Q7	6.49	-1.09	0	1
Q8	5.92	-1.12	0	1
Q9	11.06	-1.24	0	1
Q10	22.58	-1.21	0	1
Q11	6.34	-1.11	0	1
Q12	8.24	-1.12	0	1
Q13	3.54	-0.41	0	1
Q14	4.99	-0.83	0	1

2 Evaluation: Evidence of Model and Item fit

2.1 Model Fit

Model fit- Evidence of issues, somewhat mixed bag: M2 significant, but low RMSEA & TLI, CFI close to 1; suspect it relates to high correlation among items

[156]:
```
M2(fit2PL)
```

A data.frame: 1 × 9

	M2 <dbl>	df <int>	p <dbl>	RMSEA <dbl>	RMSEA_5 <dbl>	RMSEA_95 <dbl>	SRMSR <dbl>
stats	158.7335	65	8.283518e-10	0.04721049	0.03791812	0.0565313	0.04890

2.2 Item Fit

- Item fis a mixed bag as well
- smaller is better for S_X2
- p(S_X2) should be > 0.05 (or whatever alpha you preset)
- Q12 and Q6 are dubious
- Q3 couldn't be estimated due to 0 degrees of freedom, probably because of a perfect fit (everyone below a certain Θ got Q3 wrong and everyone above got it correct). To Do: Verify this [ ]

[157]:
```
itemfit(fit2PL)
```

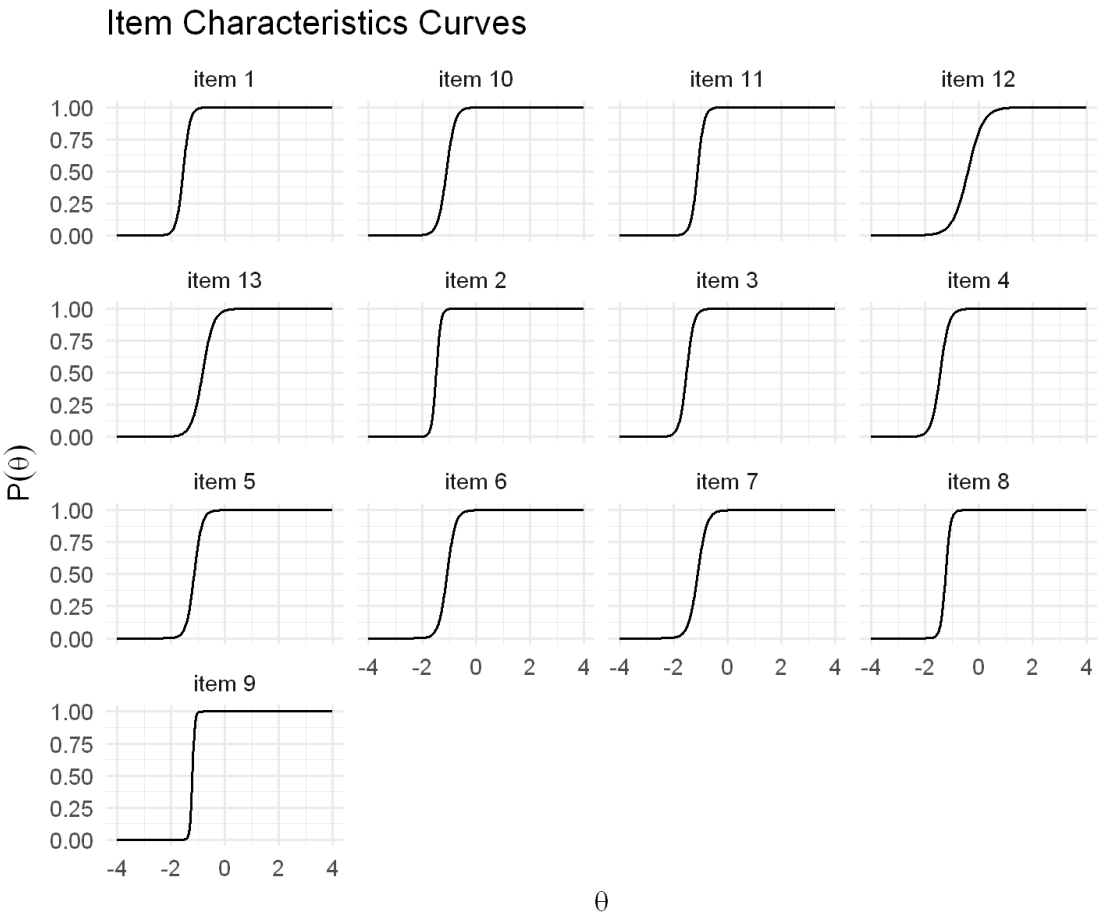
A mirt_df: 13 × 5

item <chr>	S_X2 <dbl>	df.S_X2 <dbl>	RMSEA.S_X2 <dbl>	p.S_X2 <dbl>
Q2	0.9389302	1	0.000000000	0.332553125
Q3	NaN	0	NaN	NaN
Q4	0.5645302	2	0.000000000	0.754073749
Q5	7.6091125	4	0.037343763	0.106993030
Q6	12.9000941	5	0.049417297	0.024333200
Q7	5.0581533	6	0.000000000	0.536376482
Q8	21.8915886	6	0.063981717	0.001266878
Q9	5.1460628	3	0.033251286	0.161408869
Q10	3.6751748	1	0.064301920	0.055228464
Q11	6.0868390	6	0.004729658	0.413533018
Q12	7.8216451	4	0.038427581	0.098334425
Q13	3.2961047	3	0.012351220	0.348185171
Q14	4.1589612	5	0.000000000	0.526763978

3 Visualizing Item and Test Characteristics

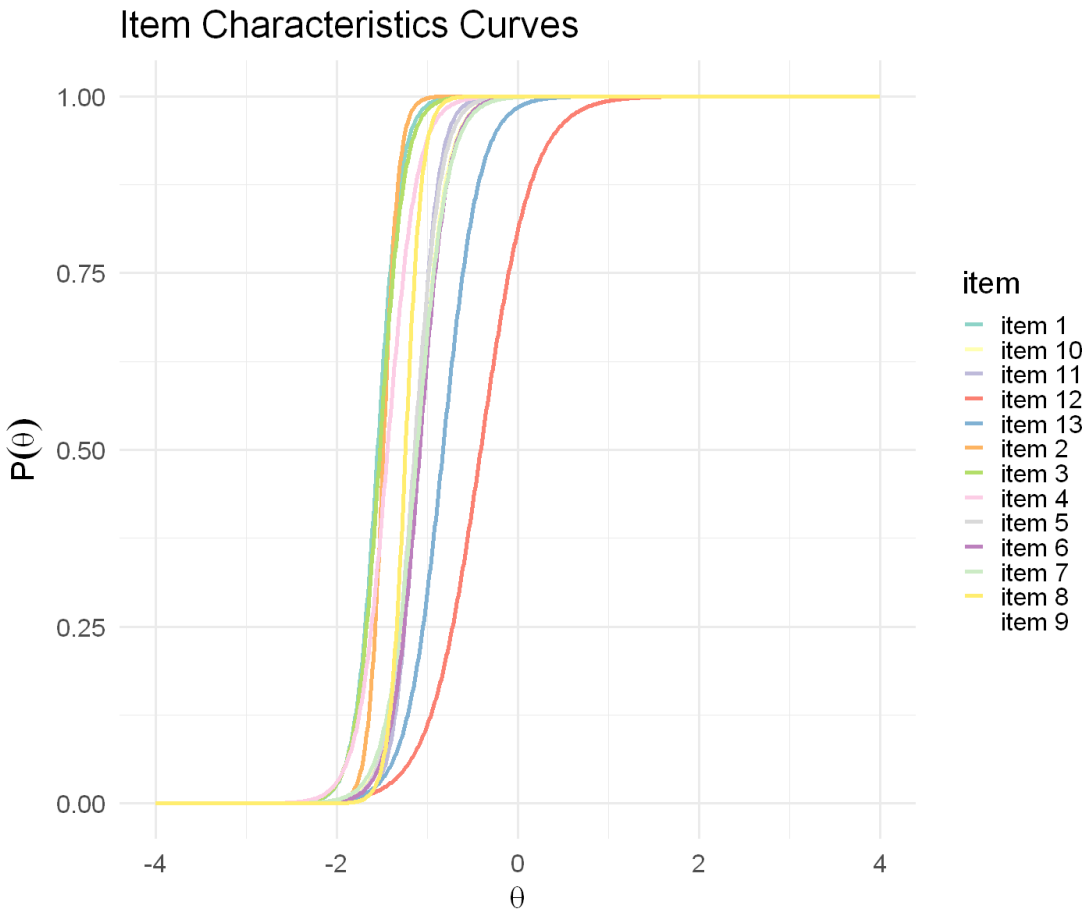
Note that on all subsequent plots Qn = Item n-1

[158]:
```
options(repr.plot.width=12, repr.plot.height=10)
tracePlot(fit2PL) + theme_minimal(base_size = 22)+ geom_line(size = 1)
```

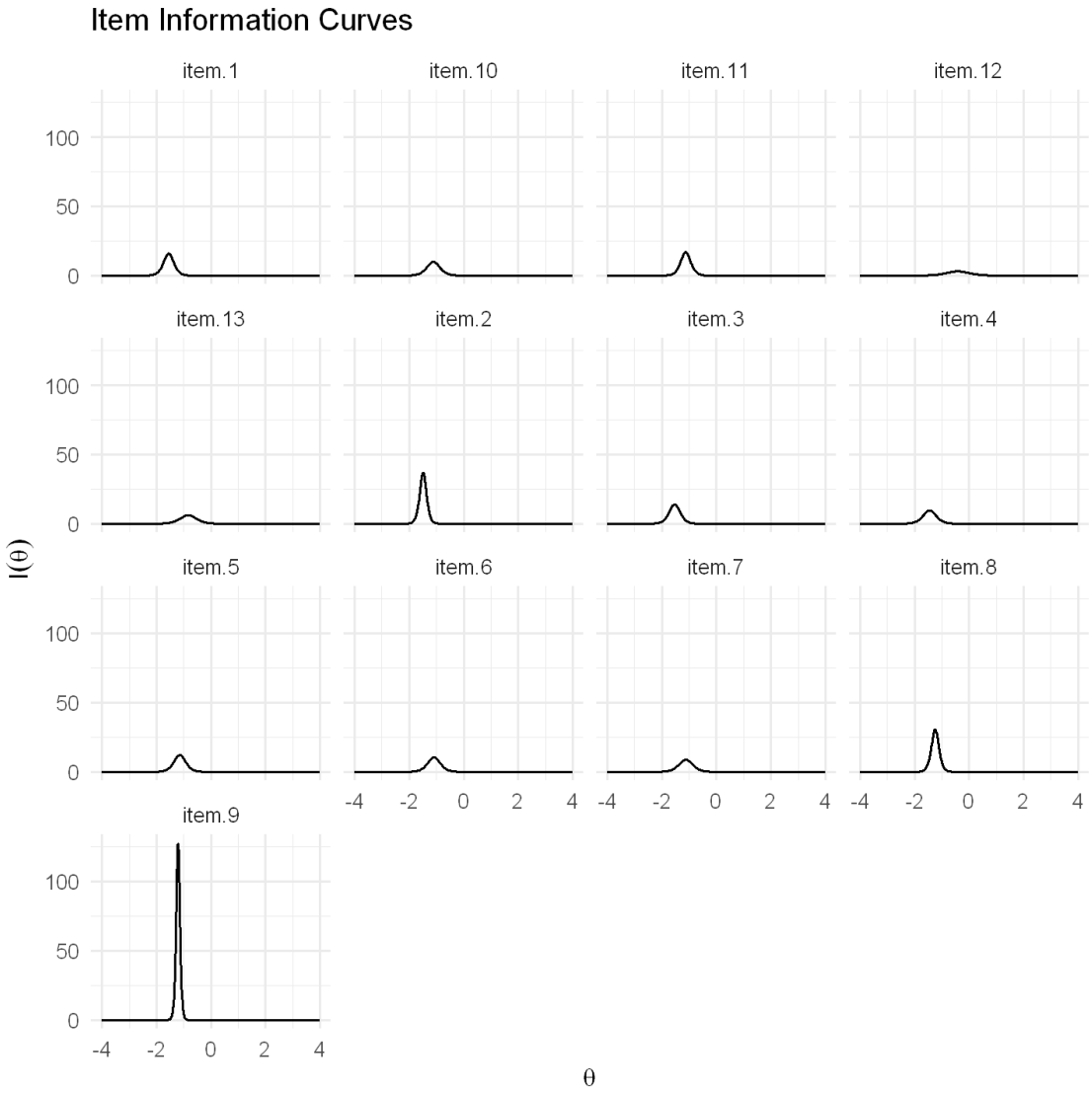


[159]:
```
tracePlot(fit2PL, facet = F, legend = T)+ scale_color_brewer(palette = "Set3")␣
↪+ theme_minimal(base_size = 24)+ geom_line(size = 1.5)
```

```
Scale for colour is already present.
Adding another scale for colour, which will replace the existing
scale.
```



[160]:
```
options(repr.plot.width=12, repr.plot.height=12)
itemInfoPlot(fit2PL, facet = T)+ theme_minimal(base_size = 20) + geom_line(size␣
↪= 1)
```



4 Information Plots

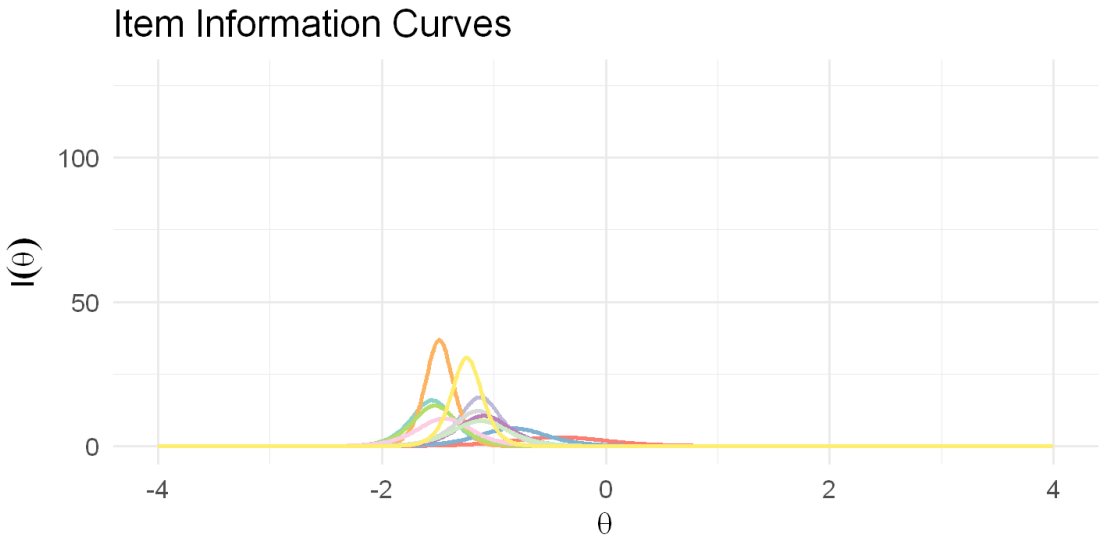
Item quality can also be expressed by representing the statistical *information* (I) an item provides. Higher information –> more accurate score estimates. The information a given question provides varies by student DQL ability (Θ)

Note again that Qn = Item n-1

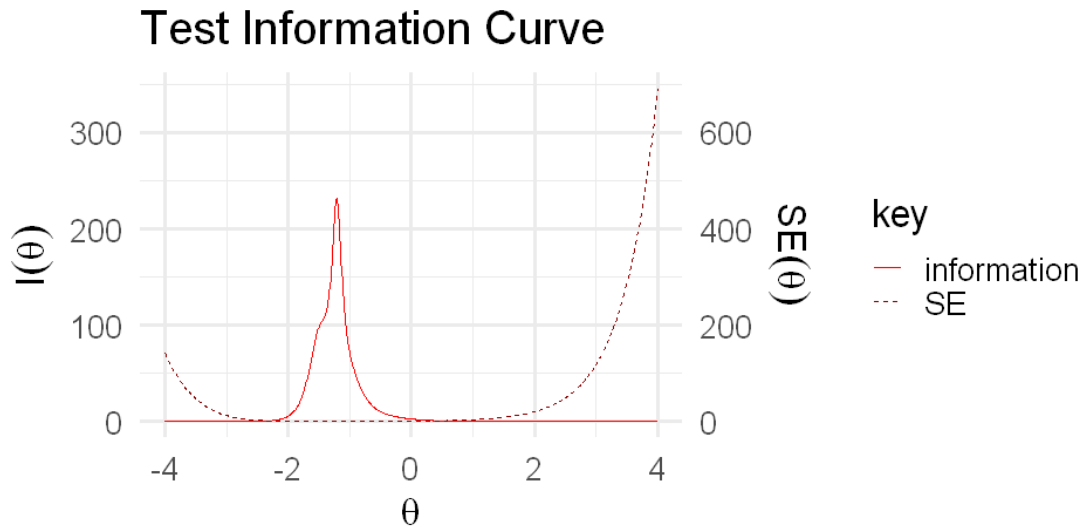
[161]:
```
options(repr.plot.width=12, repr.plot.height=6)
itemInfoPlot(fit2PL) + scale_color_brewer(palette = "Set3") +␣
  ↪theme_minimal(base_size = 24)+ geom_line(size = 1.5)
```

```
Scale for colour is already present.
Adding another scale for colour, which will replace the existing
```

scale.



[162]:
```
options(repr.plot.width=8, repr.plot.height=4)
testInfoPlot(fit2PL, adj_factor = 2) + theme_minimal(base_size = 20)
```



5 Bottom Line Takeaways

- Test appears on the whole excellent at delineating test takers, but *almost exclusively at a relatively low level of overall DQL ability*

- Test is less good at distinguishing at high levels of performance
- Per IRT, this is a reflection of both the test and its takers simultaneously
- If this is a certification test or certification test prep, where questions are externally imposed, these results are outstanding
- If this is a home-grown test, this analysis indicates a real opportunity to improve ability measurement across full spectrum of expected test taker ability
- 80/20 rule again: much of this is apparent from a histogram of test scores, albeit much more informally & intuitively

6 Areas for Future Investigation

- Explore alternative IRT models (eg Rasch)
- Comparisons to Classical Test Theory (CTT)
- Specify a multidimentional model that measures multiple latent ability factors (maybe per tags in original data)