Rust Assessment

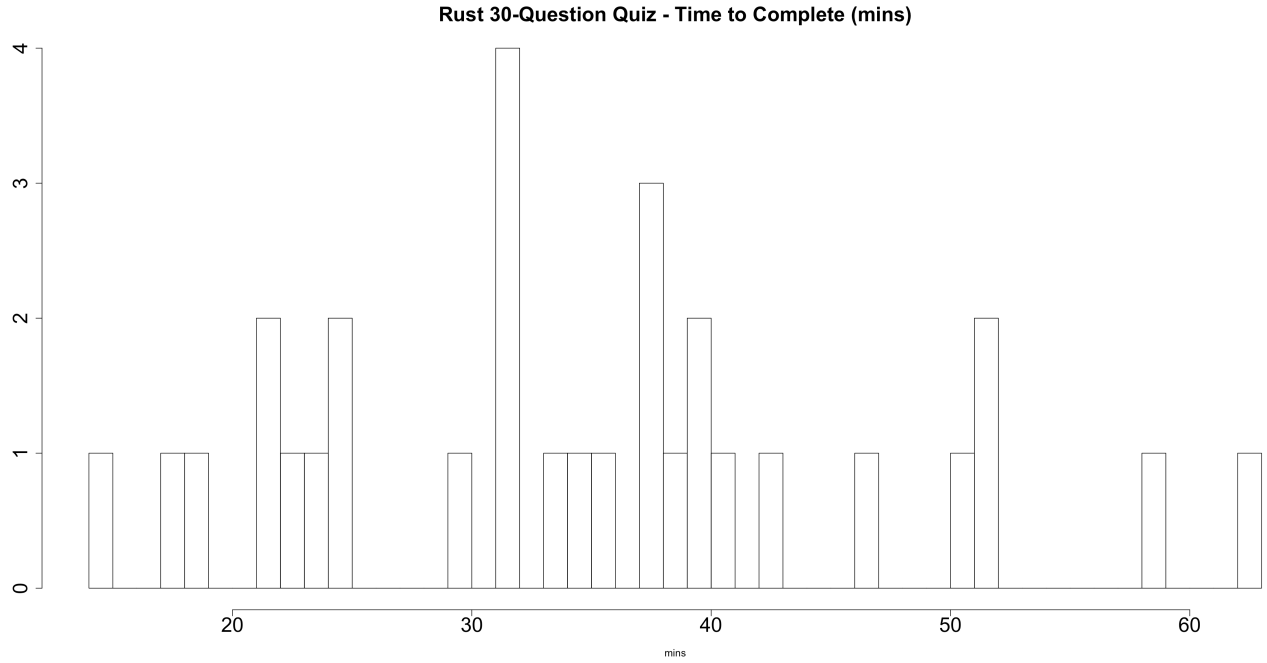
Code

Calibration, Scoring, & Impact Analyses

Number of Questions Attempted

2	6	8	30	NA
2	1	2	31	0

Time spent on Rust Quiz



n	min	q1	avg	med	q3	max
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
31	14.37	24.41	34.94065	34.8	40.18	62.43

1 row

Time spent on Quiz, by Claim

Time spent on Quiz, by Question

Analytic Sample

Exclusion Criteria:

- test/fake accounts
- attempted all questions
- total time to complete > 15min (30s per question)

The proportion of candidates who completed the Rust Quiz of those who attempted it is 0.861.

The proportion of candidates who have a valid attempted is 0.833 of those who attempted the quiz.

Classical Test Theory Analysis

Item mean is an indicator of item difficulty. Items with extremely low or high means are indicators of poorly performing items as they are overly difficult or easy, respectively, for test taskers. (parallels IRT item difficulty parameter)

Item-total correlation is the correlation between a score on a particular item and the performance on the rest of the test. High item-total correlations would indicate that test takers who score well on the overall test generally also performed well on the individual item. Low or negative item-total correlations are indicators of poor item performance as they suggest that test takers who score well on the overall test score lower on

the individual item. This kind of item performance would suggest that the item may be measuring something different than the other items on the test or the item may be keyed in the wrong direction. (parallels IRT item discrimination parameter)

Alpha-deleted is a measure of the test reliability (internal consitency) - a measure of domain sampling and the impact of a flawed items if a particular item is removed. Alpha gives an estimate of average inter-item correlation among the items. An indicator of poor item performance is if once an item is removed then the overall test reliability is greatly improved.

Items are flagged to be reviewed or removed if:

- *Item means* are outside of the range 0.20 and 0.80 - indicating really easy or difficult items
- *Item-total correlations* are less than 0.20; the larger the better - indicating may be measuring something different from the rest of the test
- *Alpha-deleted* increase if the item is removed from the test - indicating may be measuring something different from the rest of the test or is adding random noise (unsystematic variation) to the overall test.

```
Overall Reliability is: 0.878
```

df_blueprint$question_id	flag	name	target	n_candidates	item_mean	item_total_correlation	alpha_delete
1386	<=	lang_rust_01	rust-container-hashmap	30	0.800	0.216	0.87
1387	<=	lang_rust_02	rust-containers-iterator	30	0.933	0.155	0.87
1388	<=	lang_rust_03	rust-container-mut-loop	30	0.533	0.056	0.88
1389	<=	lang_rust_04	rust-containter-string	30	0.967	0.469	0.87
1390	<=	lang_rust_05	rust-containers-vec	30	0.967	0.402	0.87
1391		lang_rust_06	rust-domain-box	30	0.533	0.524	0.87
1392	<=	lang_rust_07	rust-domain-cargo	30	0.833	0.401	0.87
1393		lang_rust_08	rust-domain-mod	30	0.667	0.445	0.87
1394		lang_rust_09	rust-domain-string	30	0.733	0.740	0.86
1395	<=	lang_rust_10	rust-domain-tests	30	0.833	0.450	0.87
1396	<=	lang_rust_11	rust-features-borrowing	30	0.933	0.250	0.87
1397	<=	lang_rust_12	features-generics	30	0.833	0.142	0.88
1398		lang_rust_13	rust-features-lifetimes	30	0.700	0.797	0.86
1399	<=	lang_rust_14	rust-features-move	30	0.400	0.256	0.87
1400		lang_rust_15	rust-features-trait-bounds	30	0.567	0.564	0.87
1401		lang_rust_16	rust-libraries-futures	30	0.767	0.624	0.86
1402		lang_rust_17	rust-libraries-libc	30	0.567	0.298	0.87
1403		lang_rust_18	rust-libraries-rand	30	0.700	0.584	0.87
1404	<=	lang_rust_19	rust-libraries-serde	30	0.833	0.484	0.87
1405		lang_rust_20	rust-libraries-tokio	30	0.767	0.534	0.87
1406	<=	lang_rust_21	rust-std-format	30	0.667	0.171	0.88
1407	<=	lang_rust_22	rust-std-option	30	0.633	0.242	0.87
1408		lang_rust_23	rust-std-process	30	0.533	0.460	0.87
1409	<=	lang_rust_24	rust-std-result	30	0.933	0.638	0.87
1410		lang_rust_25	rust-std-sync	30	0.433	0.359	0.87
1411	<=	lang_rust_26	rust-syntax-characters	30	0.967	0.469	0.87
1412	<=	lang_rust_27	rust-syntax-integers	30	0.933	0.638	0.87
1413		lang_rust_28	rust-syntax-slice	30	0.667	0.287	0.87
1414	<=	lang_rust_29	rust-syntax-struct	30	0.833	0.601	0.87
1415		lang_rust_30	rust-syntax-tuple	30	0.767	0.731	0.86

IRT Analysis

Unidimensional Model

IRT Item Parameters

	a	b	g	u
1413	0.535	-0.685	0	1
1394	2.016	-0.173	0	1
1409	1.156	-1.602	0	1
1404	0.894	-1.240	0	1
1408	0.922	0.404	0	1
1386	0.481	-2.049	0	1
1412	1.156	-1.602	0	1
1399	0.526	1.267	0	1
1398	2.657	0.009	0	1
1397	0.392	-2.998	0	1
1395	0.787	-1.418	0	1
1405	1.083	-0.647	0	1
1400	1.173	0.302	0	1
1391	1.004	0.416	0	1
1402	0.572	0.091	0	1
1396	0.507	-3.493	0	1
1406	0.367	-1.182	0	1
1392	0.788	-1.416	0	1
1393	0.850	-0.294	0	1
1388	0.278	0.098	0	1
1403	1.073	-0.318	0	1
1407	0.436	-0.625	0	1
1401	1.280	-0.526	0	1
1389	0.741	-2.763	0	1
1387	0.439	-4.045	0	1
1415	1.943	-0.313	0	1
1390	0.647	-3.148	0	1
1411	0.741	-2.763	0	1
1410	0.645	0.952	0	1
1414	1.284	-0.862	0	1

Item Fit Statistics (S-X2): p-val < 0.05 <=> generally indicaties model misfit

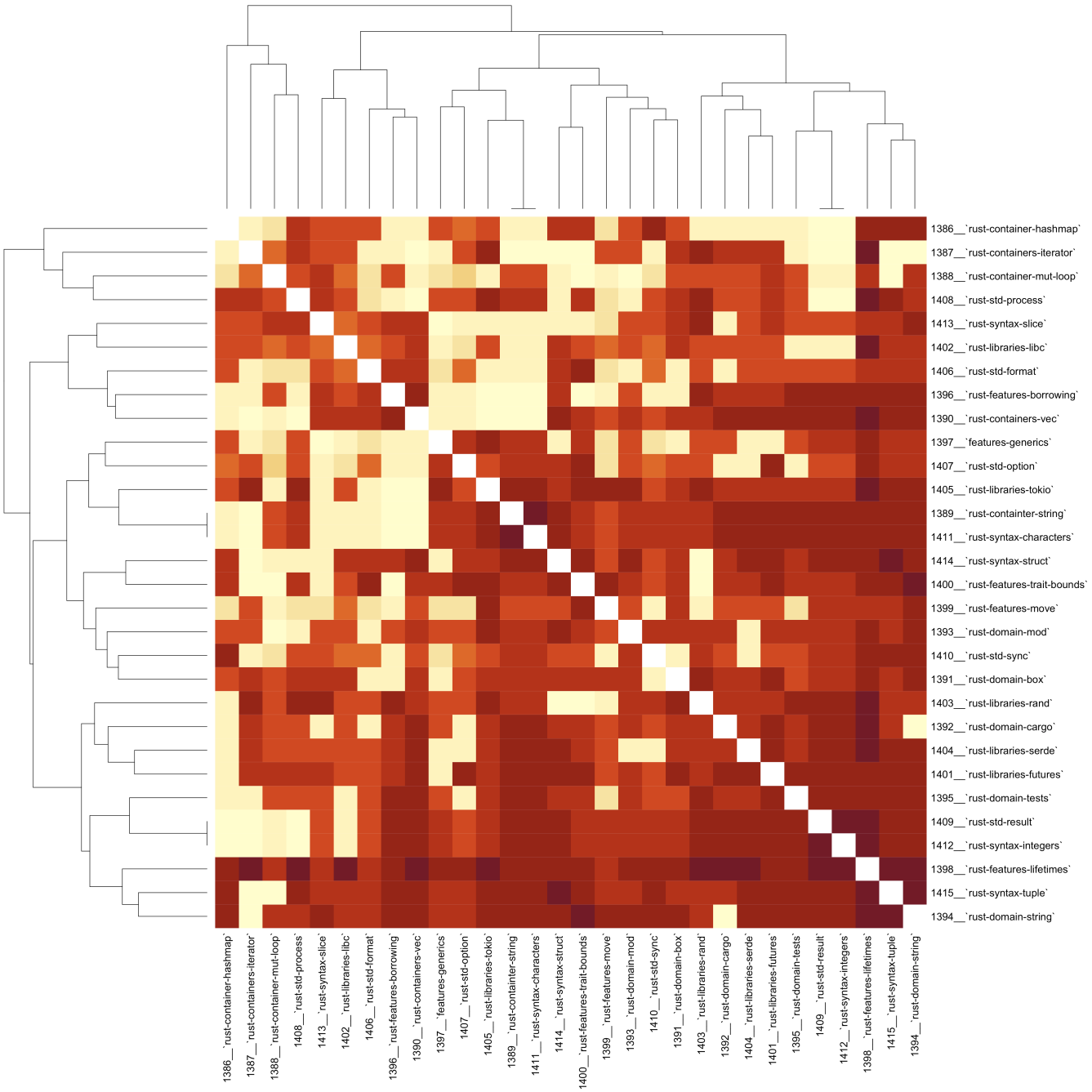
name	target	item	S_X2	df.S_X2	RMSEA.S_X2	p.S_X2	flag
lang_rust_28	rust-syntax-slice	1413	8.620	3	0.254	0.035	<=
lang_rust_09	rust-domain-string	1394	NaN	0	NaN	NaN	NA
lang_rust_24	rust-std-result	1409	NA	NA	NA	NA	NA
lang_rust_19	rust-libraries-serde	1404	NaN	0	NaN	NaN	NA
lang_rust_23	rust-std-process	1408	3.646	3	0.086	0.302	

name	target	item	S_X2	df.S_X2	RMSEA.S_X2	p.S_X2	flag
lang_rust_01	rust-container-hashmap	1386	3.763	1	0.309	0.052	
lang_rust_27	rust-syntax-integers	1412	NA	NA	NA	NA	NA
lang_rust_14	rust-features-move	1399	3.581	4	0.000	0.466	
lang_rust_13	rust-features-lifetimes	1398	NaN	0	NaN	NaN	NA
lang_rust_12	features-generics	1397	0.354	1	0.000	0.552	
lang_rust_10	rust-domain-tests	1395	0.094	1	0.000	0.759	
lang_rust_20	rust-libraries-tokio	1405	5.092	2	0.231	0.078	
lang_rust_15	rust-features-trait-bounds	1400	3.471	2	0.159	0.176	
lang_rust_06	rust-domain-box	1391	3.929	3	0.103	0.269	
lang_rust_17	rust-libraries-libc	1402	2.299	3	0.000	0.513	
lang_rust_11	rust-features-borrowing	1396	NA	NA	NA	NA	NA
lang_rust_21	rust-std-format	1406	1.040	3	0.000	0.791	
lang_rust_07	rust-domain-cargo	1392	2.292	1	0.211	0.130	
lang_rust_08	rust-domain-mod	1393	0.370	1	0.000	0.543	
lang_rust_03	rust-container-mut-loop	1388	6.678	5	0.108	0.246	
lang_rust_18	rust-libraries-rand	1403	0.949	1	0.000	0.330	
lang_rust_22	rust-std-option	1407	12.754	4	0.275	0.013	<=
lang_rust_16	rust-libraries-futures	1401	0.311	1	0.000	0.577	
lang_rust_04	rust-containter-string	1389	NA	NA	NA	NA	NA
lang_rust_02	rust-containers-iterator	1387	NA	NA	NA	NA	NA
lang_rust_30	rust-syntax-tuple	1415	NaN	0	NaN	NaN	NA
lang_rust_05	rust-containers-vec	1390	NA	NA	NA	NA	NA
lang_rust_26	rust-syntax-characters	1411	NA	NA	NA	NA	NA
lang_rust_25	rust-std-sync	1410	3.158	3	0.043	0.368	
lang_rust_29	rust-syntax-struct	1414	NaN	0	NaN	NaN	NA

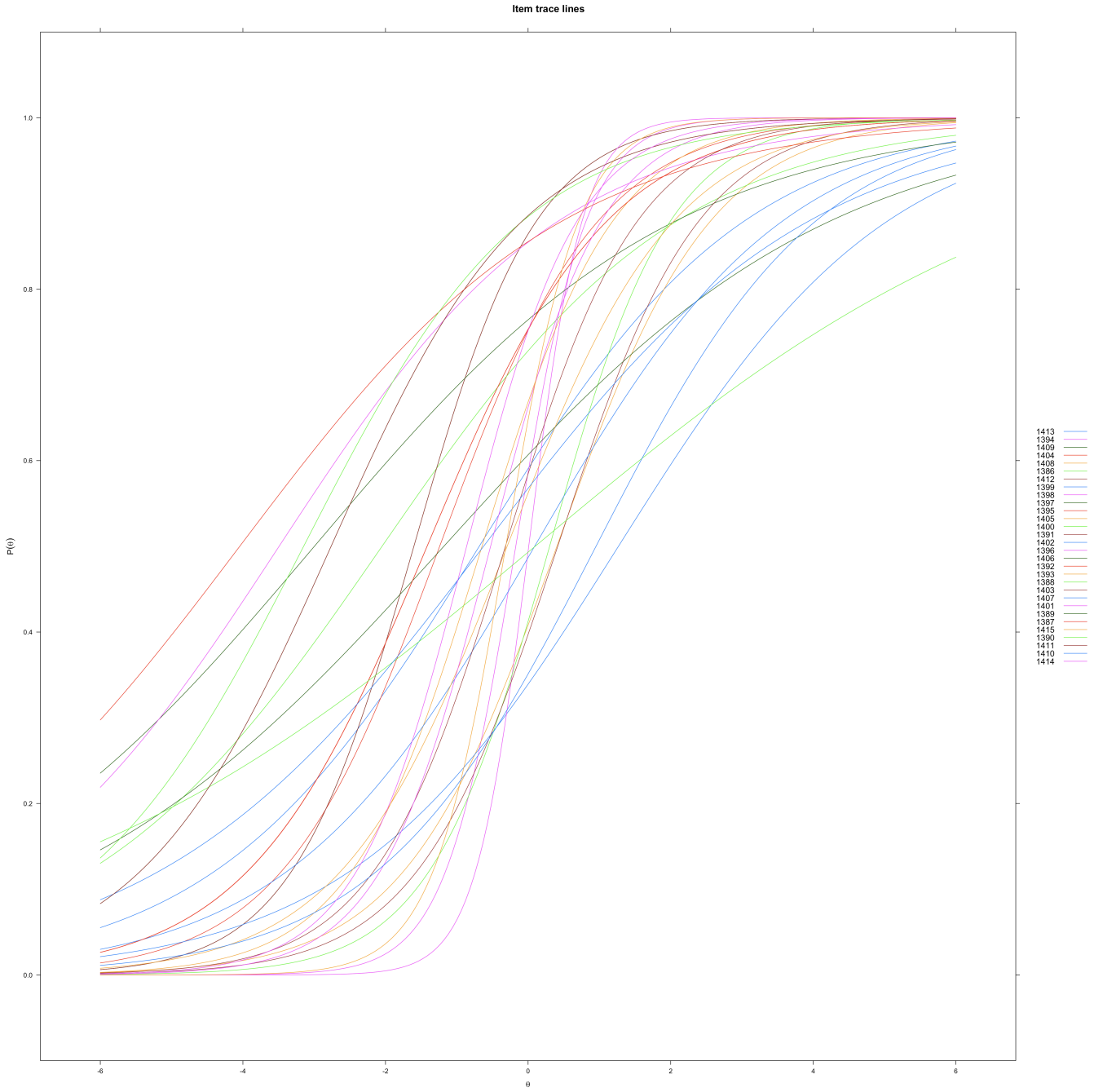
Items are flagged with '<=' to be reviewed or removed if *p-value* < 0.05. This indicates poor fit of 2PL model.

IRT Plots

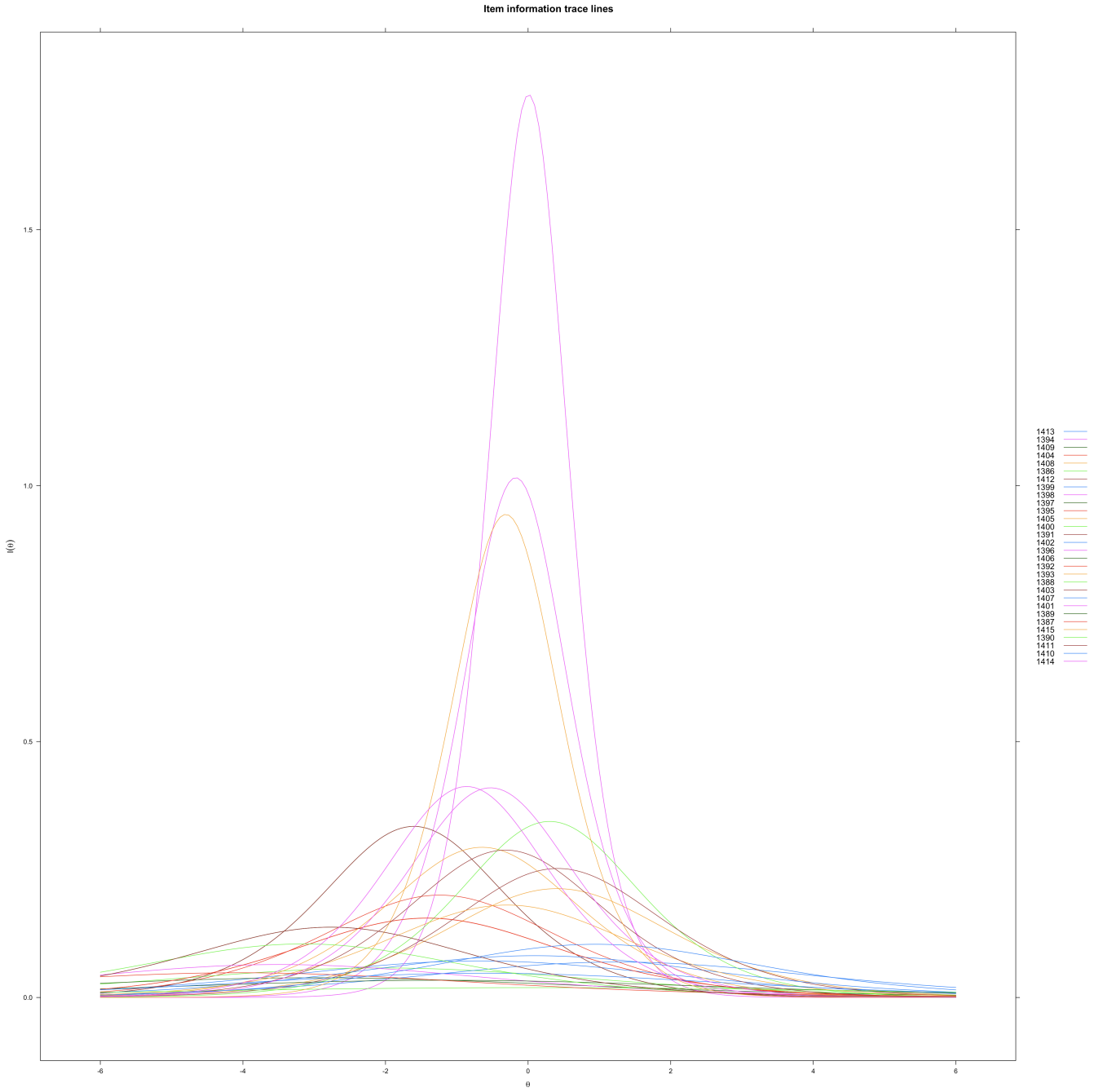
Item Local-Dependence Plots - Residual Dependencies given a unidimensional model:



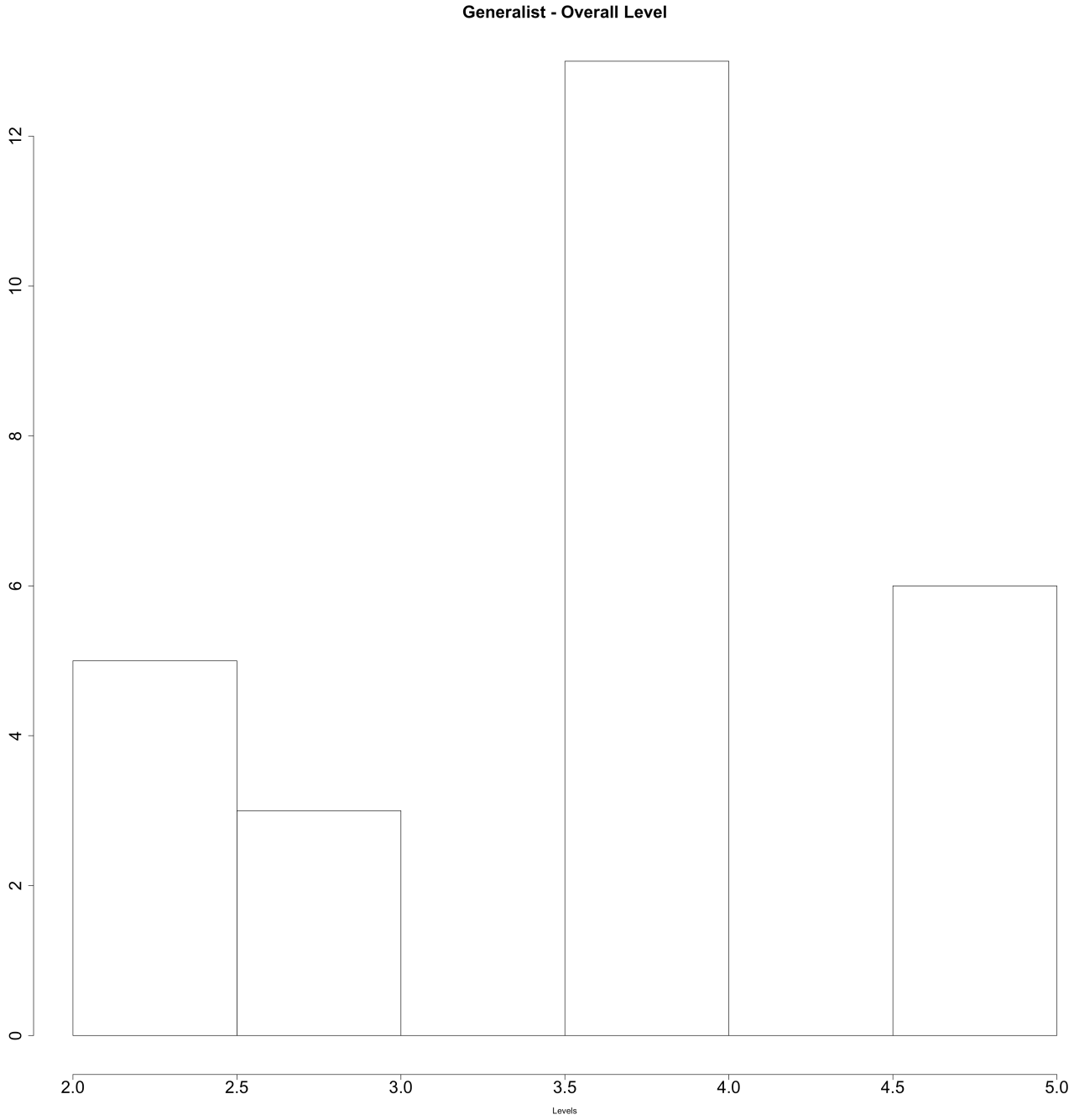
Item Characteristic Curves (Tracelines), by Claims

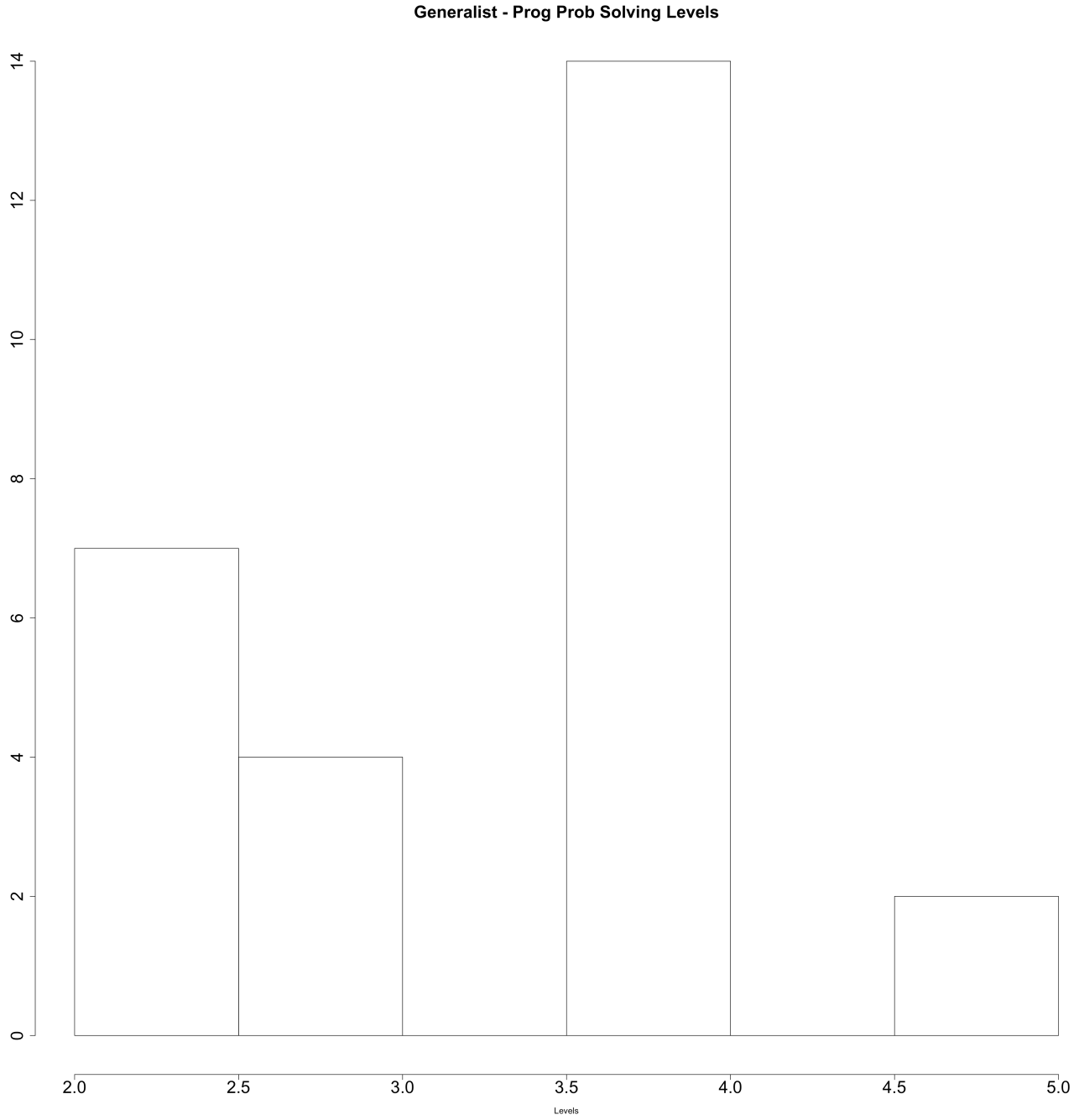


Item Information Curves by Claims



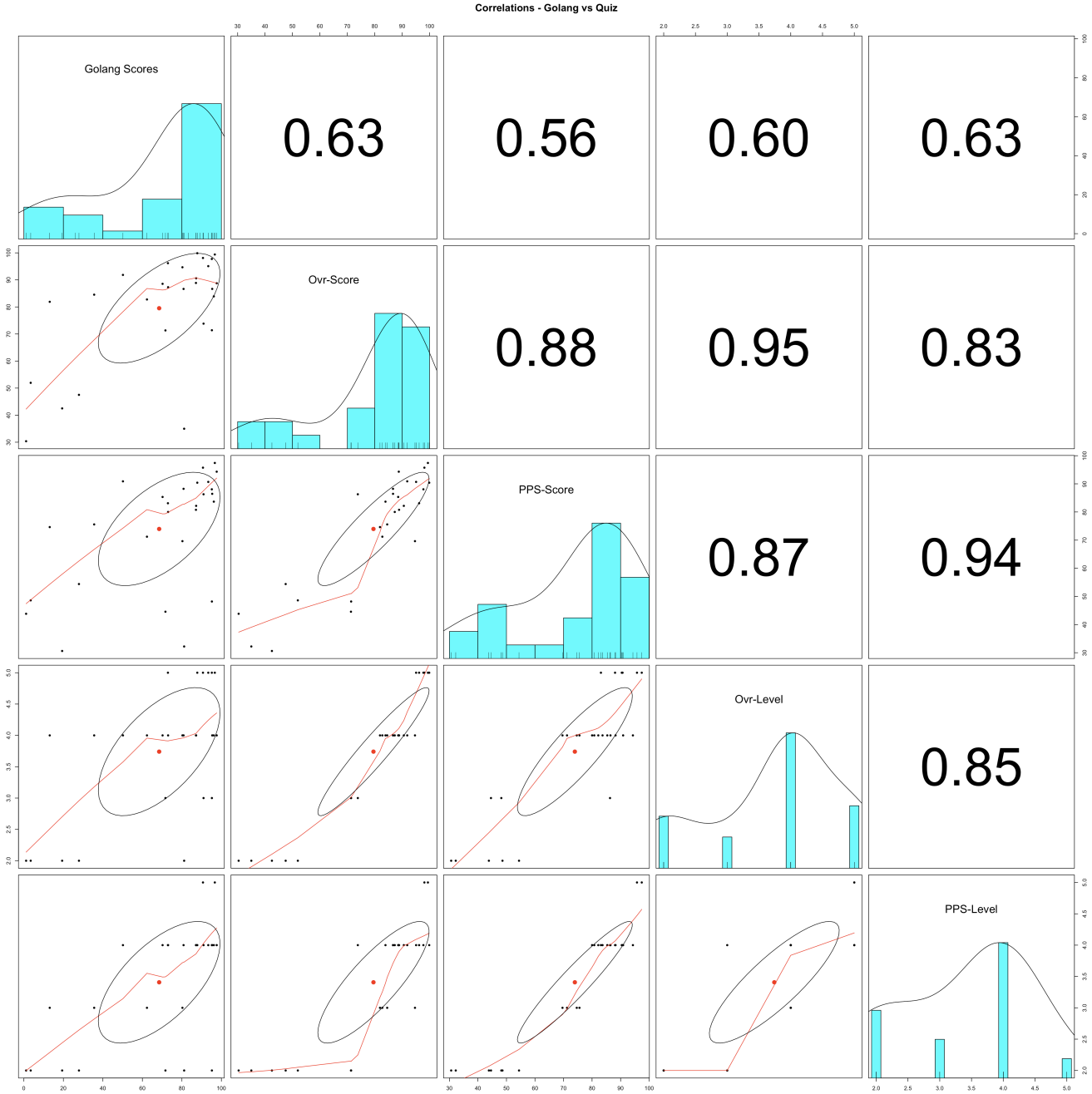
Calibration Sample Characteristics


Generalist - Overall Level
0
2
4
6
8
10
12
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Levels

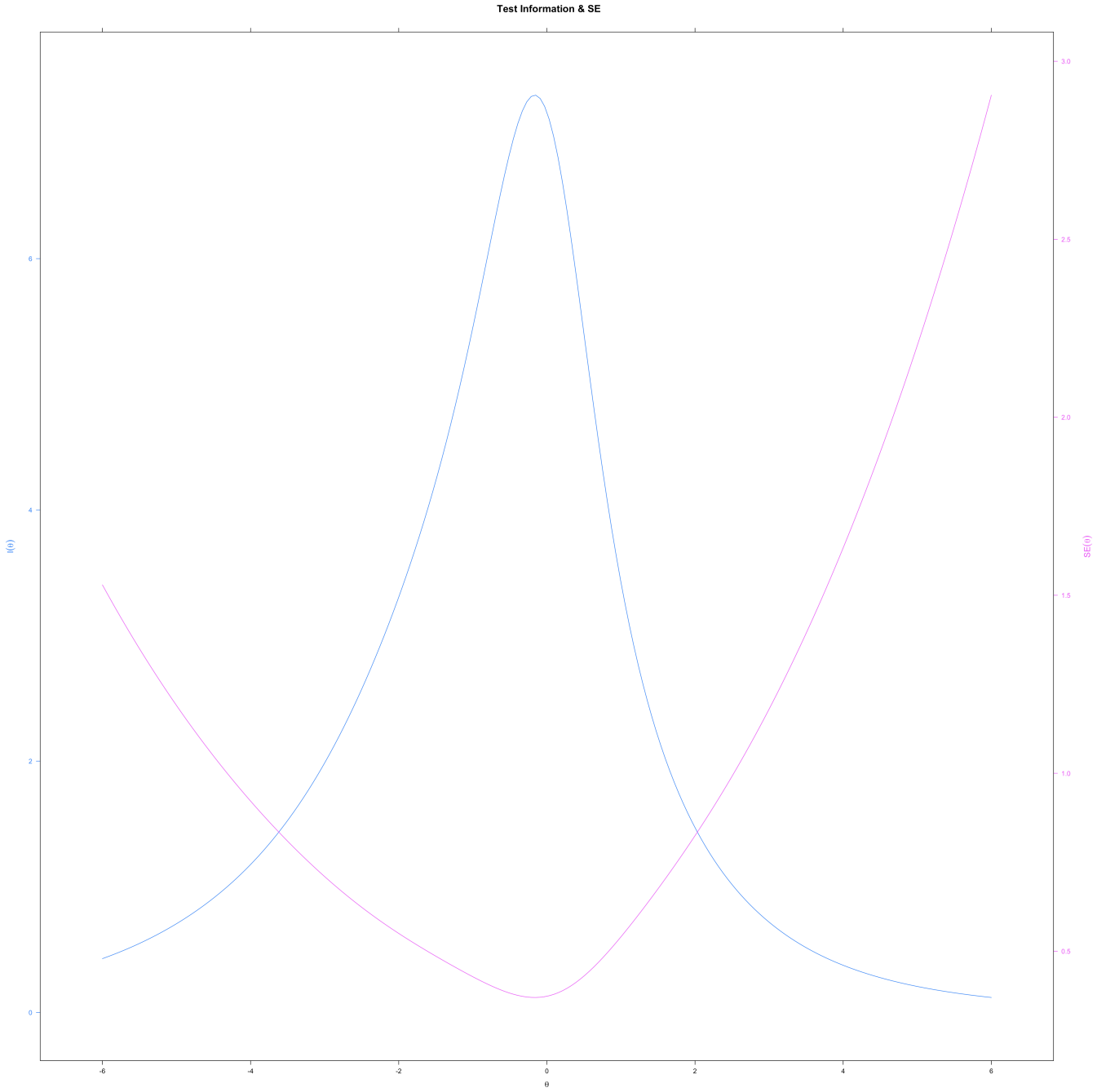


Scoring*

Distributions & Correlations


Correlations - Golang vs Quiz
Golang Scores
0.63
0.56
0.60
0.63
Ovr-Score
0.88
0.95
0.83
PPS-Score
0.87
0.94
Ovr-Level
0.85
PPS-Level

Avg. SE



Equating/Linking Scores

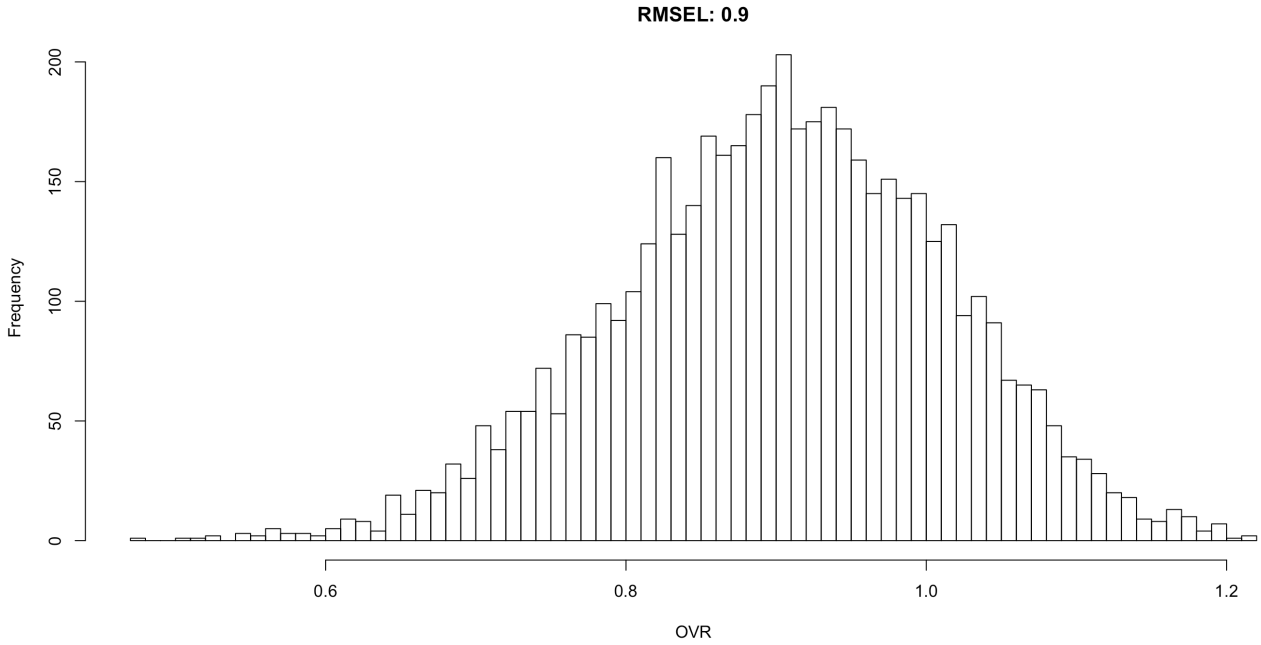
In order to get langauge assessment SkillEstimates on the same scale as the the other quizzes, it requires being able to produce item parameters or IRT-ability estimates on the same scale as the other assessments.

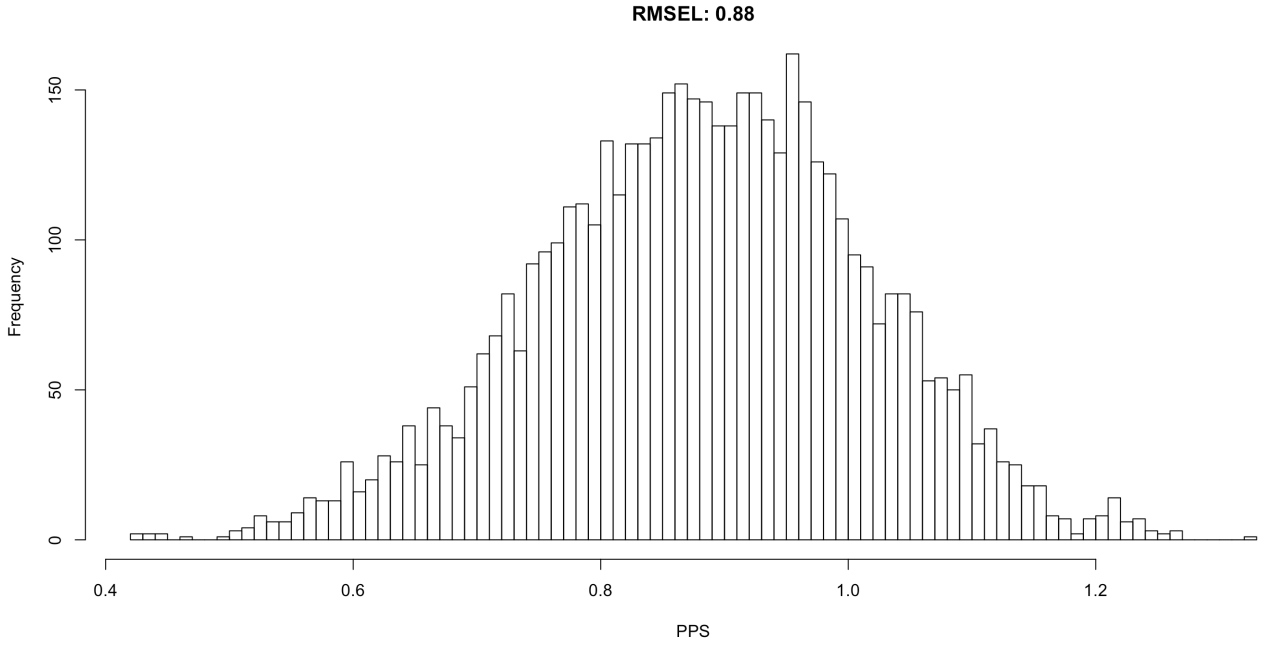
The equating/linking design available is a *Single Group* Design where the same candidates have taken both a previous core quiz and a language assessement. Therefore, there are limited methodolgies available to link the two scales together (plus with a limited sample size).

Three approaches explored are:

- Mean Method
- Linear Method
- Equipercentile Method

Mean Method





$OVR

avg	sd
<dbl>	<dbl>
0.9036407	0.1098552

1 row

$PPS

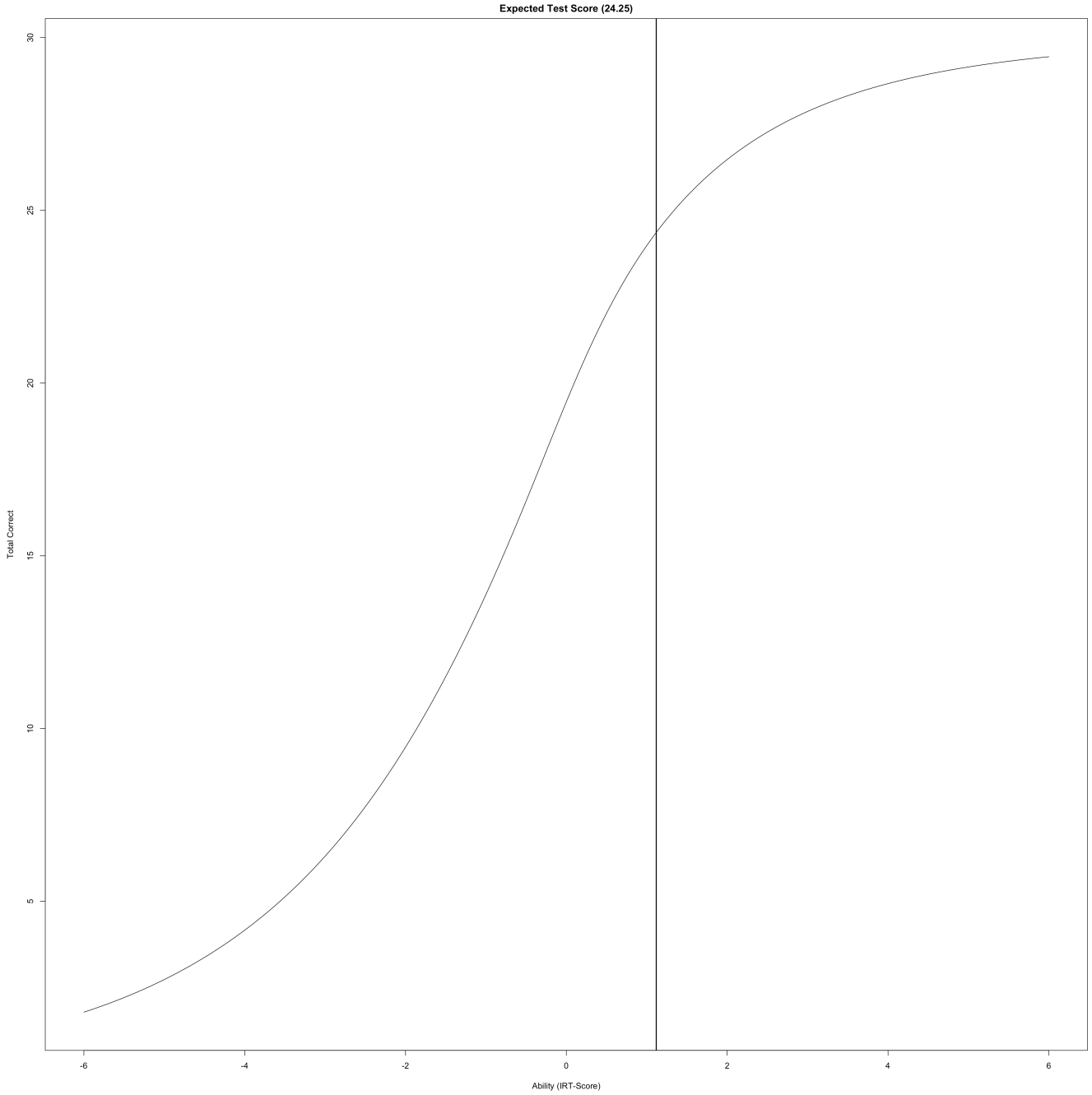
avg	sd
<dbl>	<dbl>
0.8827884	0.1330331

1 row

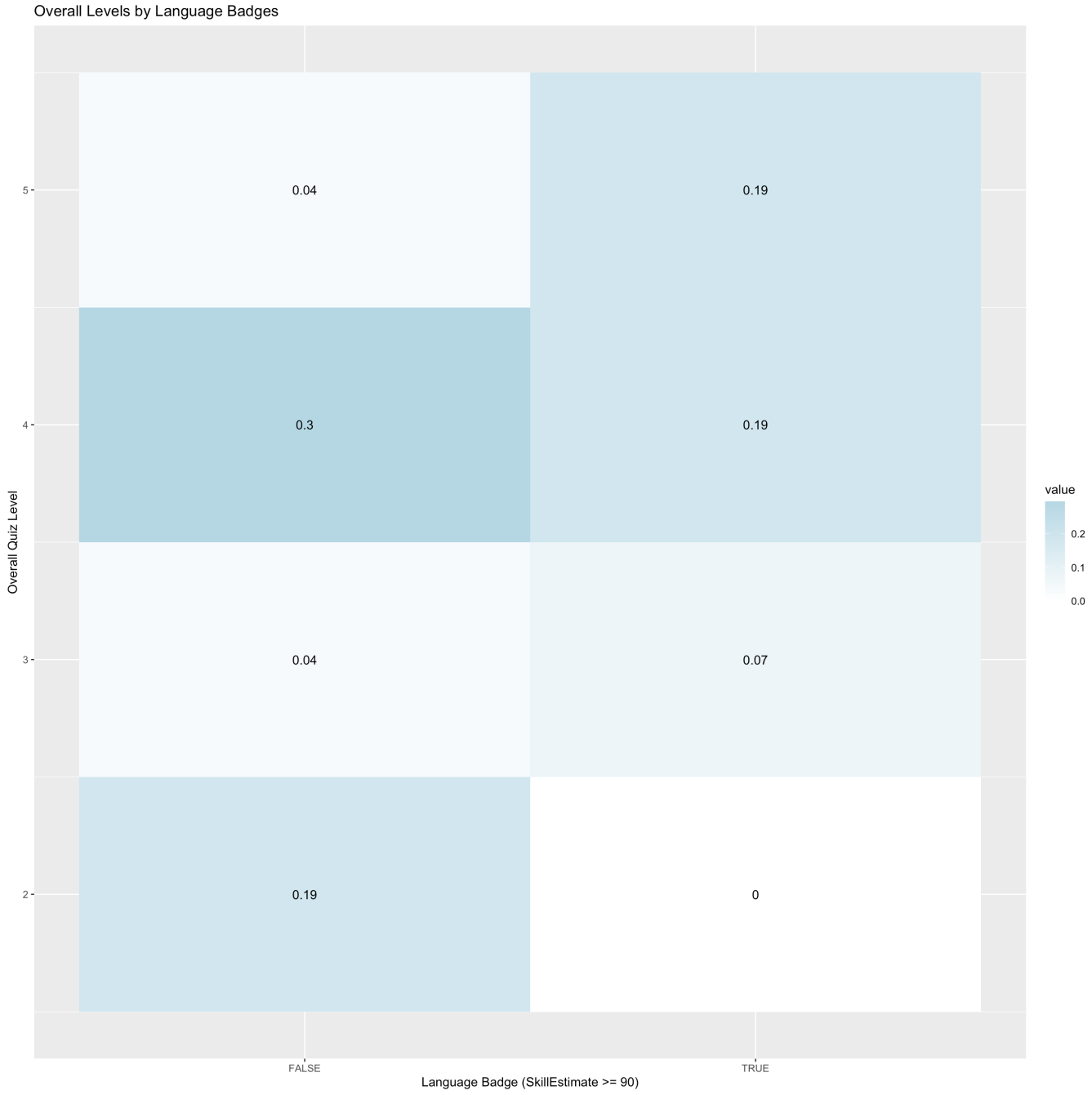
```
NA
```

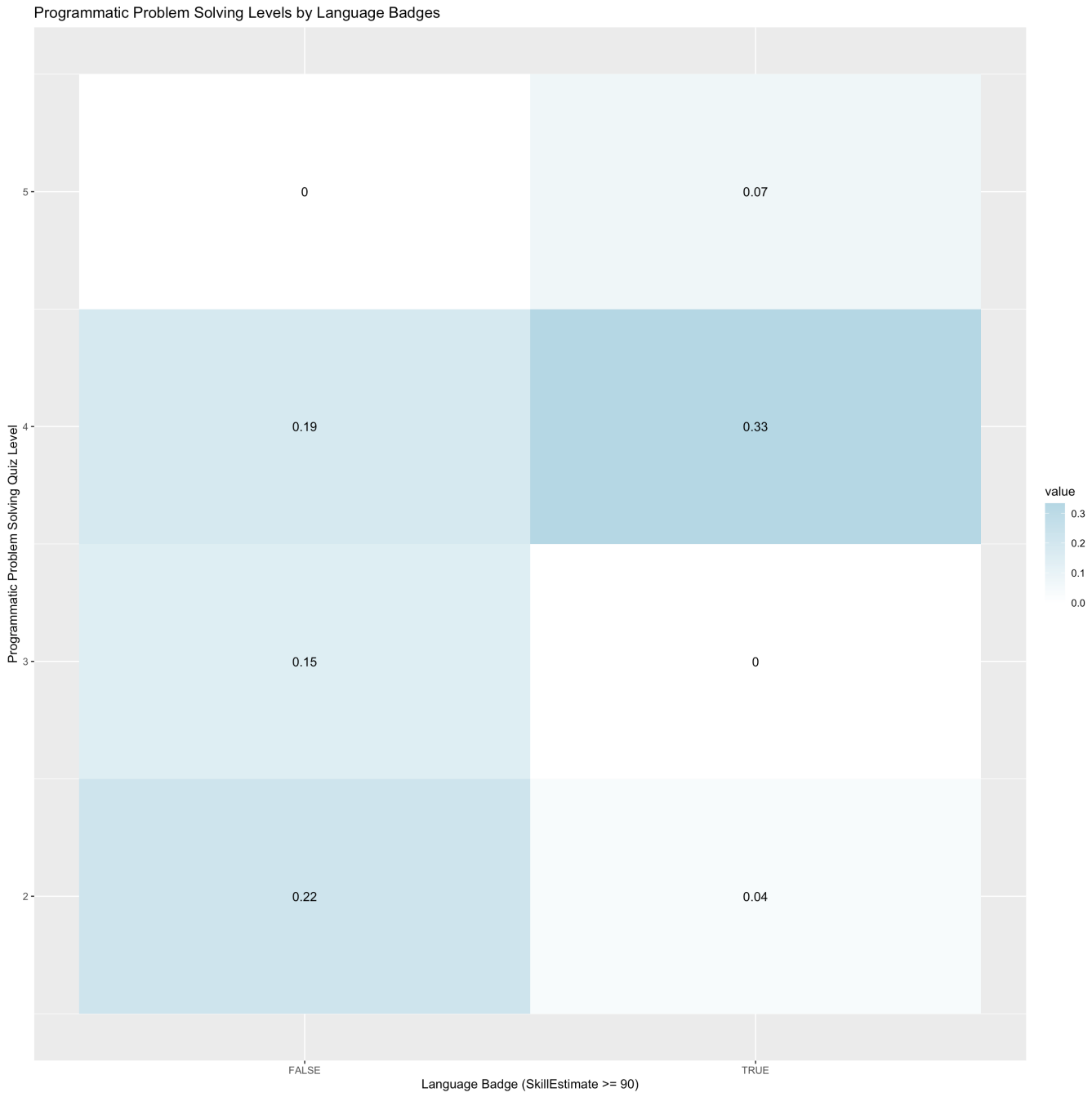
Result: Programmatic Problem Solving scale scores produce smaller linking error

Badges - Test Characteristic Curve

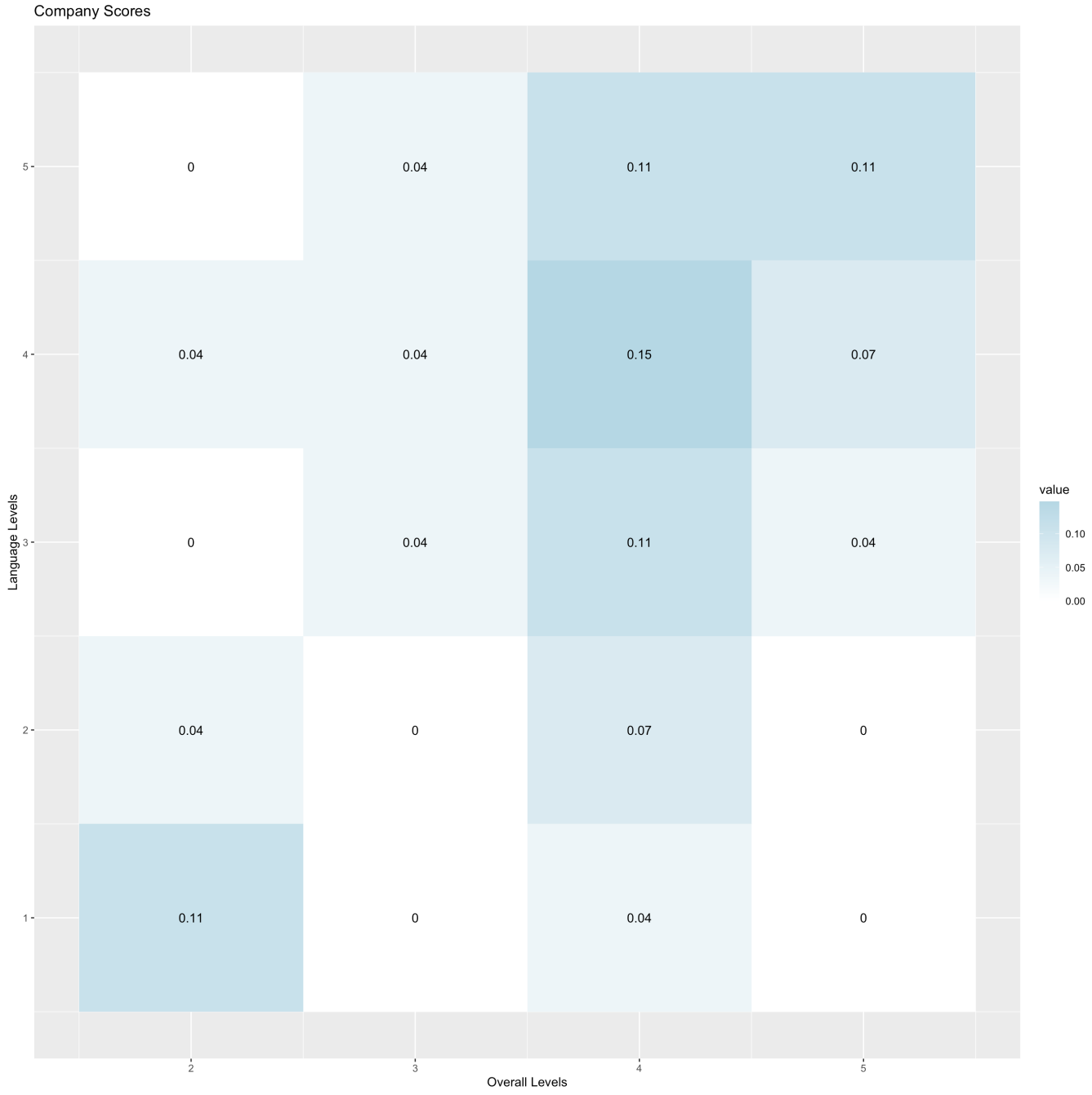


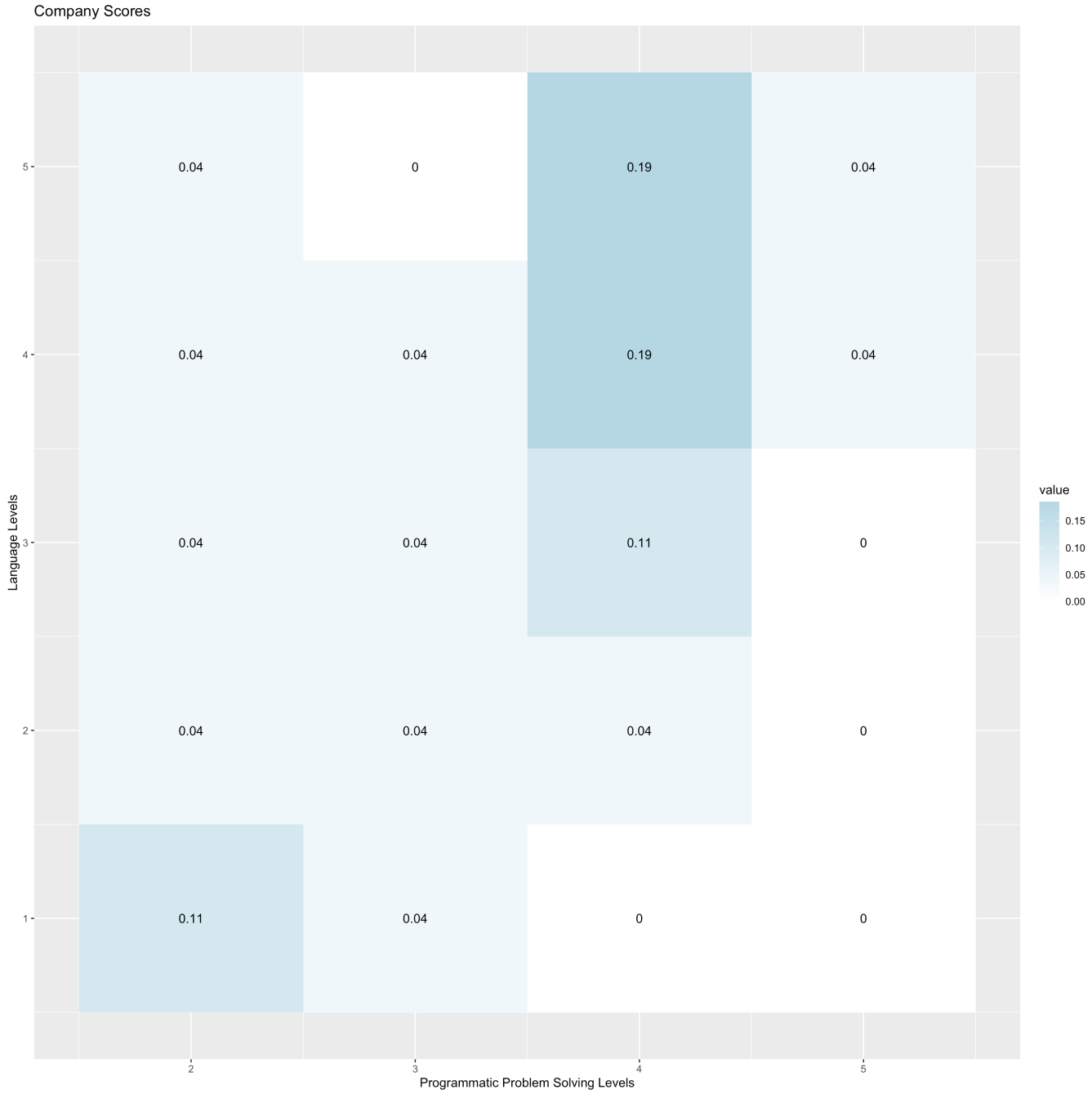
Badges - Quiz Level, by Badge


Overall Levels by Language Badges
5
4
3
2
0.04
0.19
0.3
0.19
0.04
0.07
0.19
0
Overall Quiz Level
FALSE
TRUE
Language Badge (SkillEstimate >= 90)
value
0.2
0.1
0.0

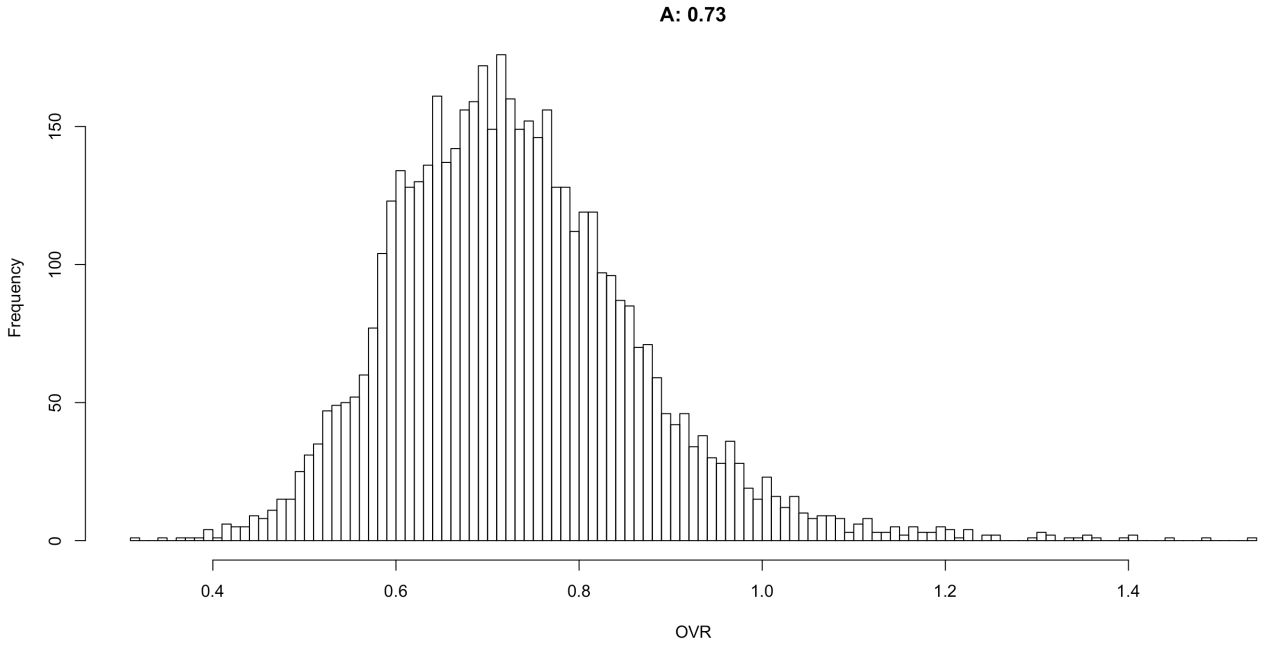


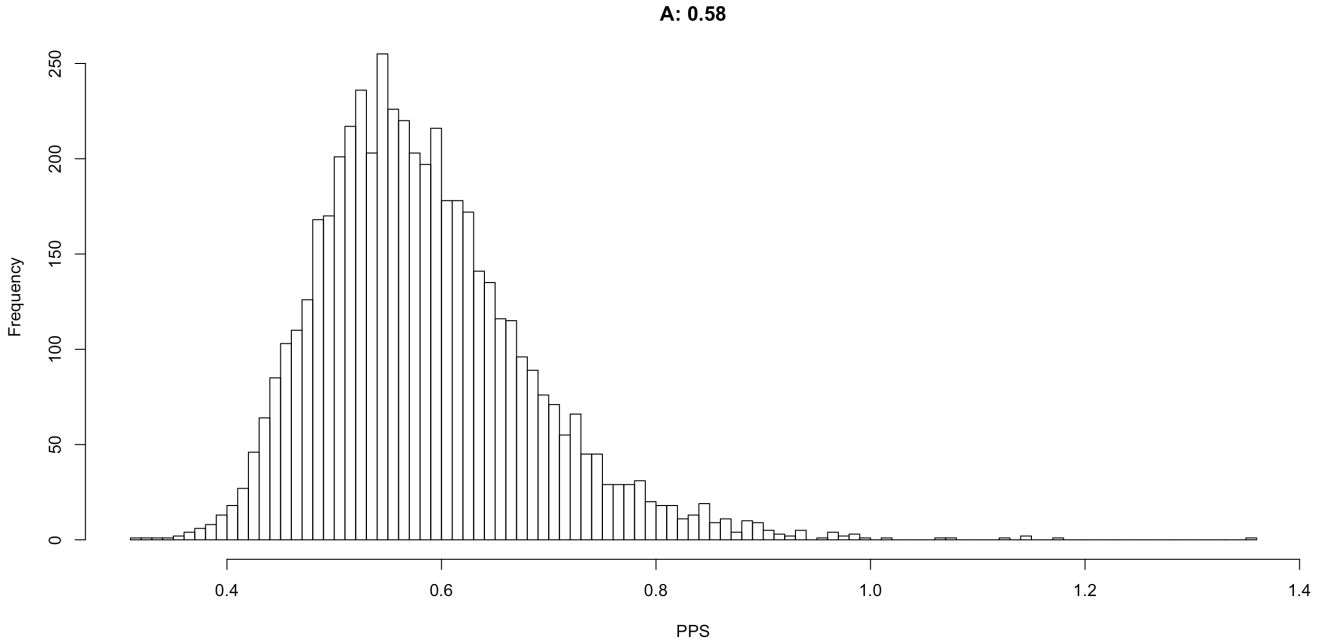
Language Levels, Quiz Levels

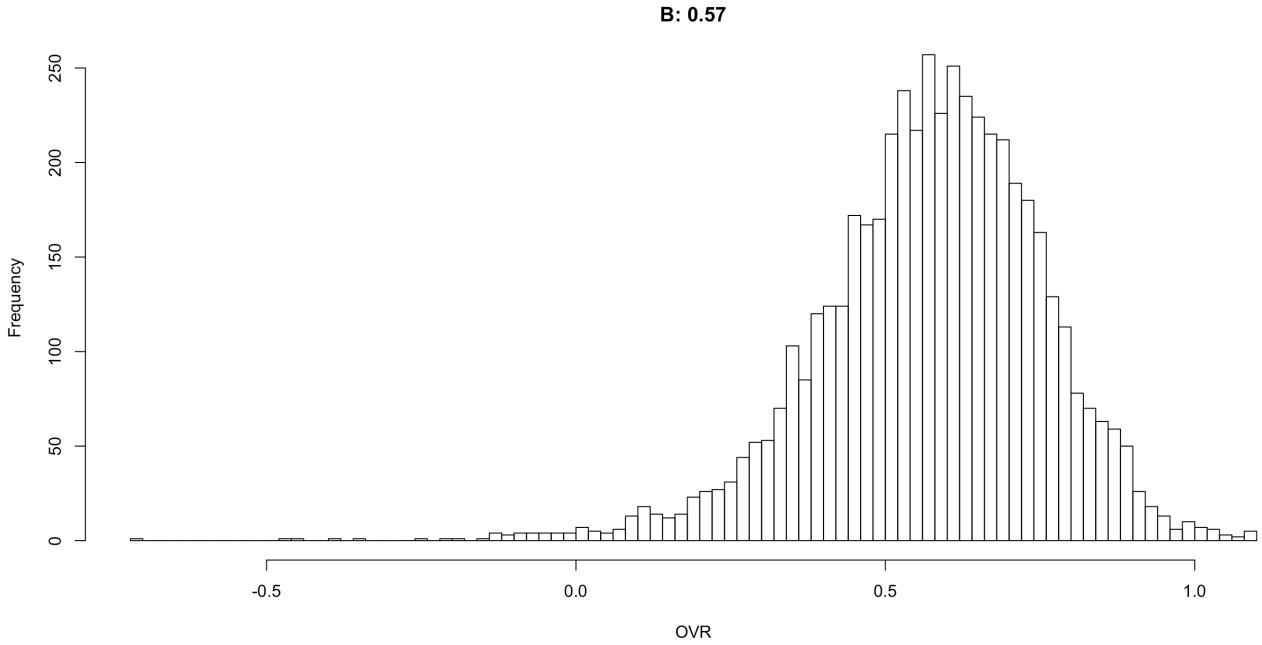

Company Scores
Language Levels
Overall Levels
value
0.10
0.05
0.00
5
4
3
2
1
2
3
4
5
0
0.04
0.11
0.11
0.04
0.04
0.15
0.07
0
0.04
0.11
0.04
0.04
0
0.07
0
0.11
0
0.04
0

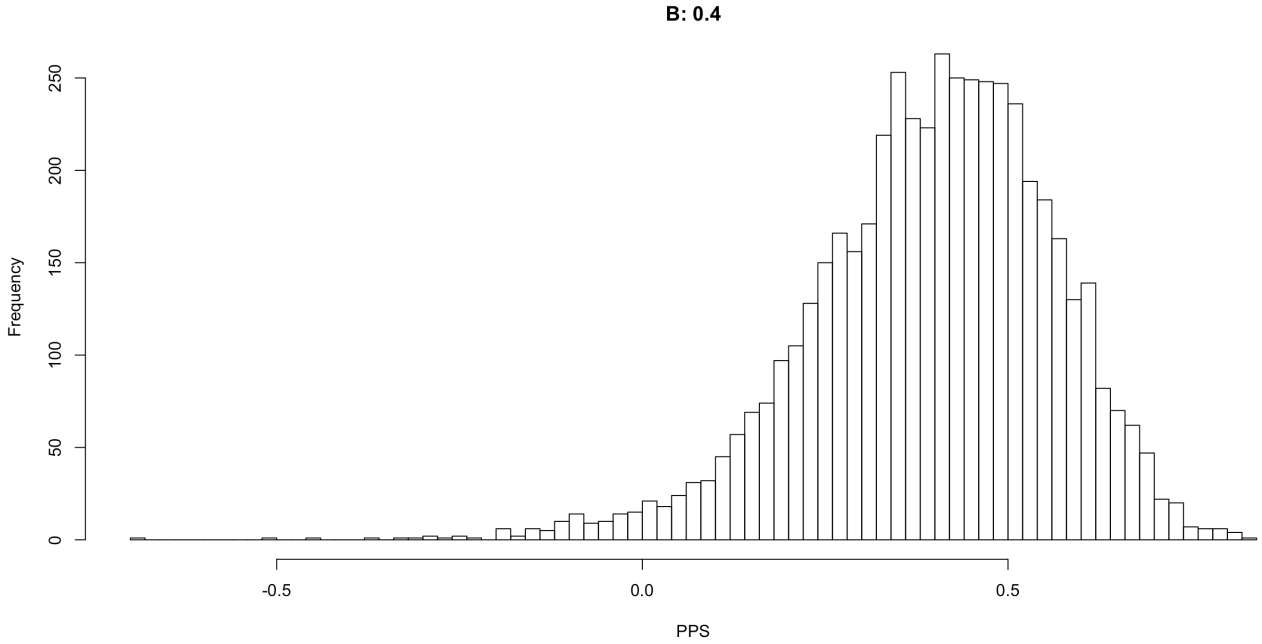


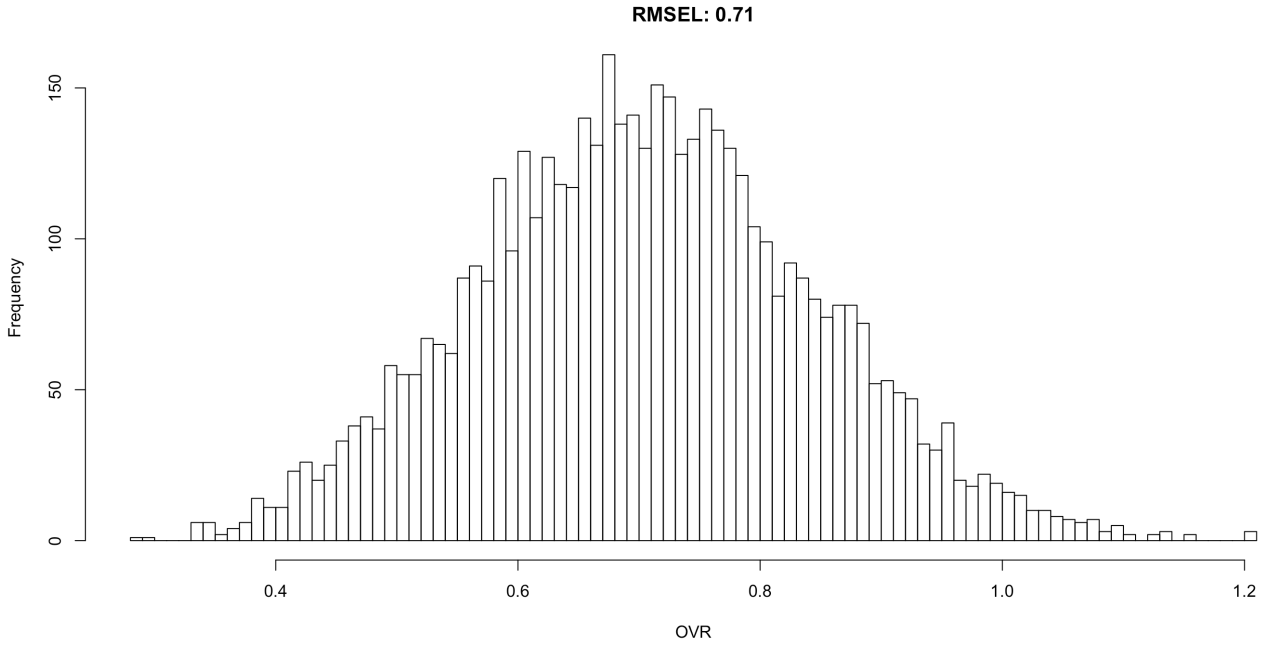
Linear Method

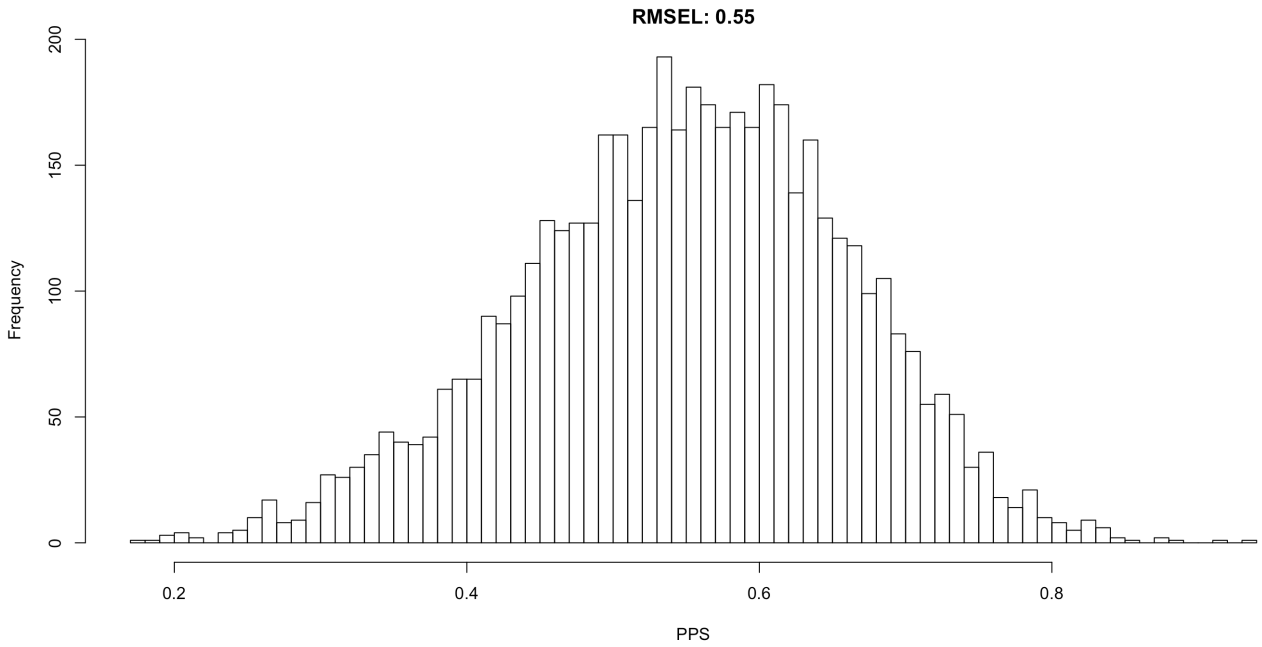

A: 0.73
Frequency
0
50
100
150
0.4
0.6
0.8
1.0
1.2
1.4
OVR


A: 0.58
Frequency
0
50
100
150
200
250
0.4
0.6
0.8
1.0
1.2
1.4
PPS


B: 0.57
Frequency
0
50
100
150
200
250
-0.5
0.0
0.5
1.0
OVR

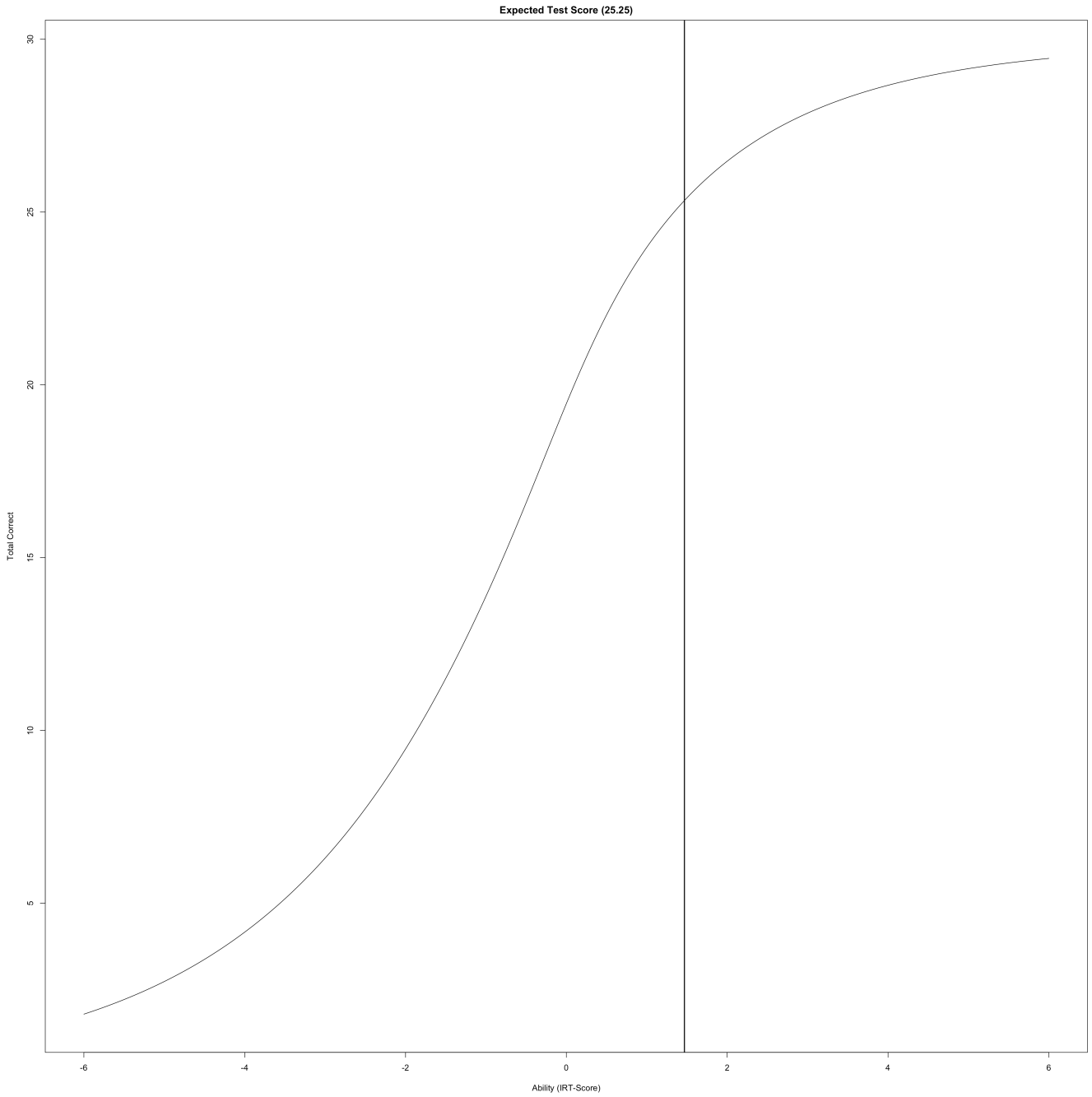

B: 0.4
Frequency
0
50
100
150
200
250
-0.5
0.0
0.5
PPS



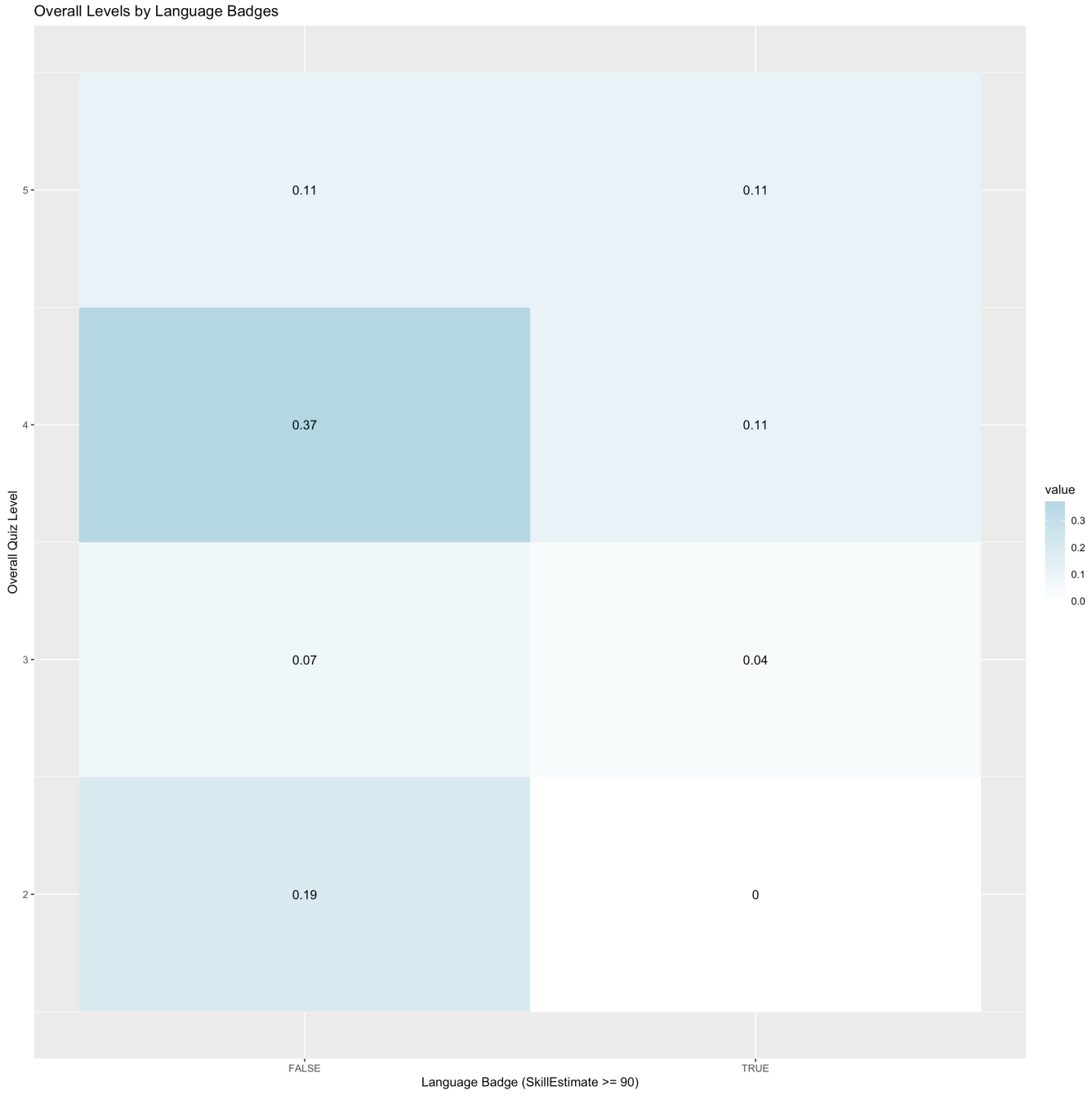


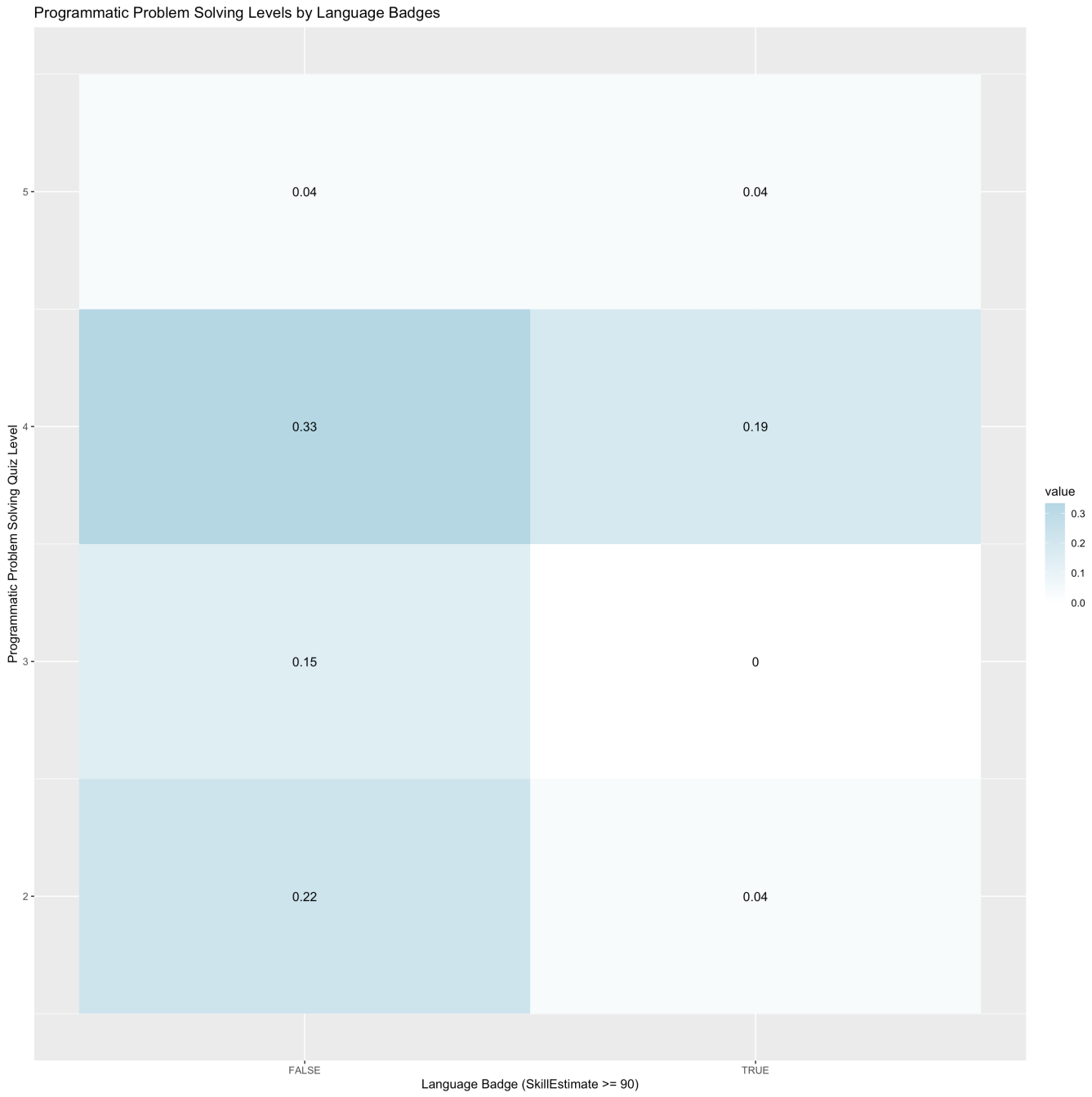
Result: Programmatic Problem Solving scale scores produce smaller linking error

Badges - Test Characteristic Curve

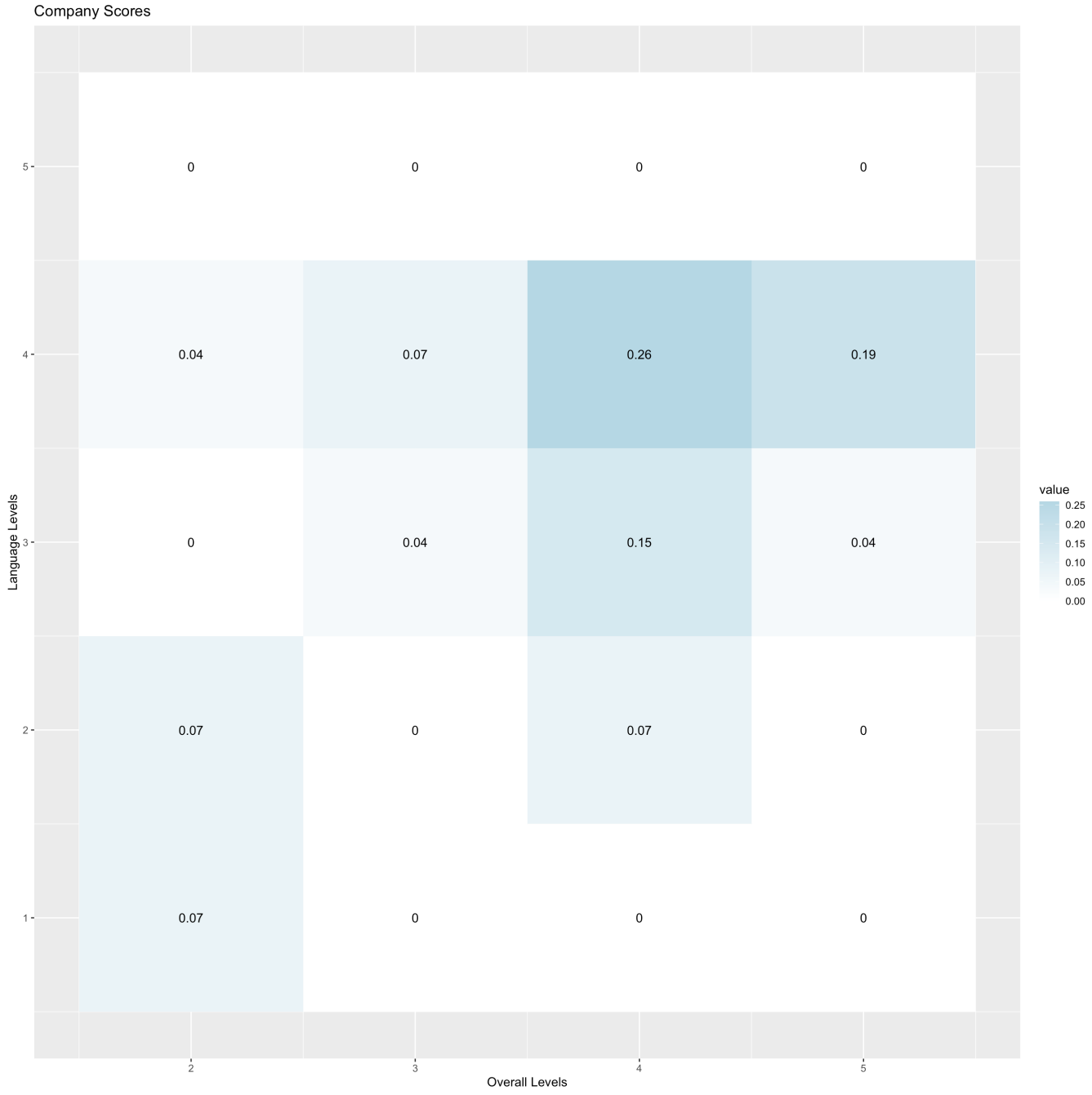


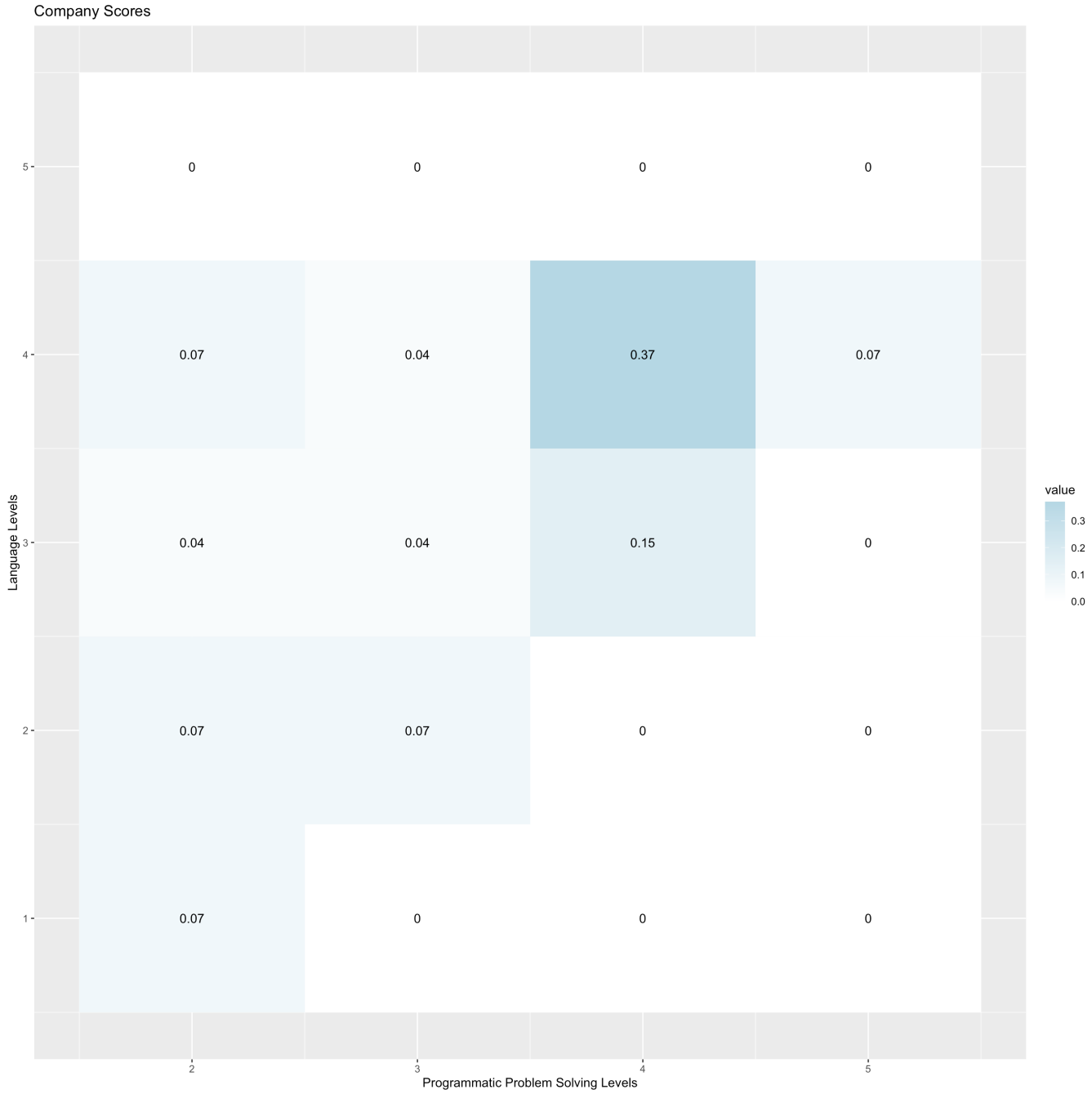
Badges - Quiz Level, by Badge


Overall Levels by Language Badges
Overall Quiz Level
5
4
3
2
0.11
0.11
0.37
0.11
0.07
0.04
0.19
0
FALSE
TRUE
Language Badge (SkillEstimate >= 90)
value
0.3
0.2
0.1
0.0



Language Levels, Quiz Levels


Company Scores
Language Levels
Overall Levels
value
0.25
0.20
0.15
0.10
0.05
0.00
5
4
3
2
1
2
3
4
5
0
0
0
0
0.04
0.07
0.26
0.19
0
0.04
0.15
0.04
0.07
0
0.07
0
0.07
0
0
0



Percentile Method

Summary