ALWAYS LEARNING

Enabling skills: Increased machine score information by deviating from human scoring model

Optimizing Raw Score Usage to Reduce Measurement Error

John H.A.L. de Jong

9th Annual UK Rasch Users Group
London March 20, 2015

Research Context

PTE Academic: 20 item types, reflecting different modes of language use, different response tasks and different response formats.

Reports scores on Pearson's ***Global Scale of English (GSE)*** from 10 to 90.

Includes 11 scores on the score report:

- an Overall Score
- four Communicative Skills scores
- six Enabling Skills scores

SEM on Overall and Communicative Skill Scores

PRESESSIONAL

DIRECT ENTRY

	Average Measurement Error				
	A2	**B1**	**B2**	**C1**	**C2**
Overall	2.5	2.4	2.7	3.2	3.5
Listening	3.7	3.4	3.8	4.4	4.9
Reading	3.9	4.0	4.4	5.2	5.8
Speaking	3.6	3.9	4.4	5.1	5.6
Writing	4.3	3.7	4.1	4.8	5.3

SEM on Enabling Skill Scores

	Average Measurement Error				
Enabling skills	**A2**	**B1**	**B2**	**C1**	**C2**
Grammar	20.7	21.6	20.5	18.7	17.8
Fluency	6.5	6.1	6.0	6.1	6.3
Pronunciation	6.4	6.5	6.3	6.3	6.4
Spelling	18.2	18.7	14.9	14.5	15.7
Vocabulary	10.9	10.7	10.8	11.4	12.3
Written discourse	28.5	29.6	28.1	26.6	26.6

Research Background

Initial machine scoring of the enabling skill scores was modelled on human ratings.

Machine scores are computed as a continuous variable based on the relative probability of a given score.

For example,

A response that has a high probability of receiving a score of 3 would be calculated as 2.95 or 3.05, while a response with equal probability of being a 2 or a 3 would be calculated as 2.5.

Automated scoring machine trained on human raters

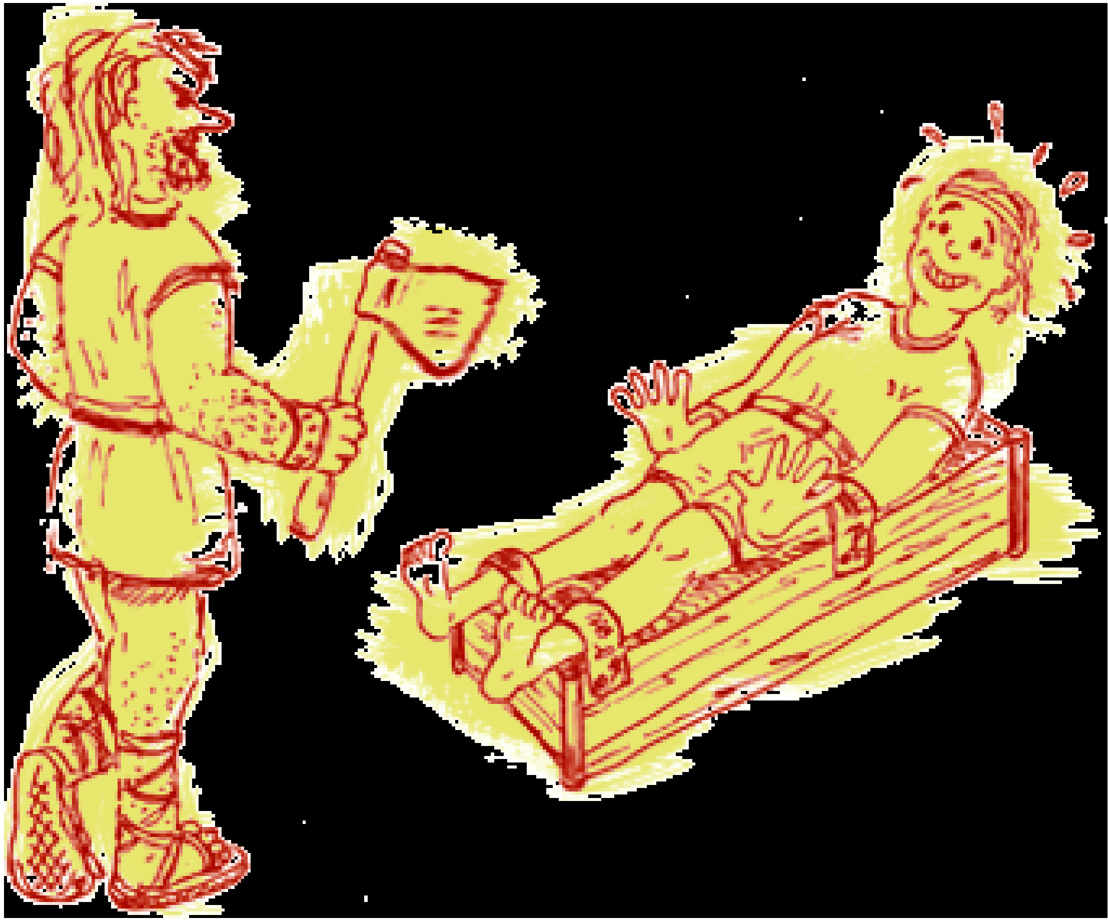
The human marking rubrics present an integer **3-point scale** (0, 1, 2)

Automated scoring algorithms are developed by training the scoring model on several hundred responses scored by human raters.

The original outcome of the trained models is continuous: a real number theoretically between $-\infty$ and $+\infty$, but in practice >95% of the estimates are limited to a range of -1 to 3.

Procrustes

In Greek mythology, **Procrustes** (Προκροὐστης) or ‘the stretcher’, was a rogue smith and bandit from Attica who physically attacked people by stretching them or cutting off their legs, so as to make them fit the size of his iron bed. In general, when something is Procrustean, different lengths or sizes or properties are fitted to an arbitrary standard.



Machine scores procrustesized

The machine scores are transformed into integers to conform to the human rater scoring rubrics and to enable IRT scaling under standard IRT models.

The original machine scores are converted:

- If score is greater than the max, set to the max (2)
- If score is less than the min, set to the min (0)
- In other cases, round to 0, 1, and 2.

WHY?

How to improve measurement precision

This research reports on the results from a project aimed at improving the precision of the scoring model for three of the Enabling skills:

- ***Written Discourse***
- ***Grammar***
- ***Vocabulary***

Two analysis phases

Phase I

Changing current machine estimates transformation

Phase II

Two step improvement

1. Changing current machine estimates transformation
2. Adding information from new traits

Analysis Phase I: Improving *Written Discourse*

- ***Written Discourse*** score is measured on two traits from one item type (Essay).
  - Development, Structure & Coherence (DSC)
  - General Linguistic Range (GLR)

Approach

Recode the original data following the new coding scheme:

- **<.5 =0**
- **.5 thru 1 = 1**
- **1 thru 1.5 =2**
- **>1.5 = 3**

This way, the data will be on a **4 point scale**, which allows a finer breakdown of test takers' scores.

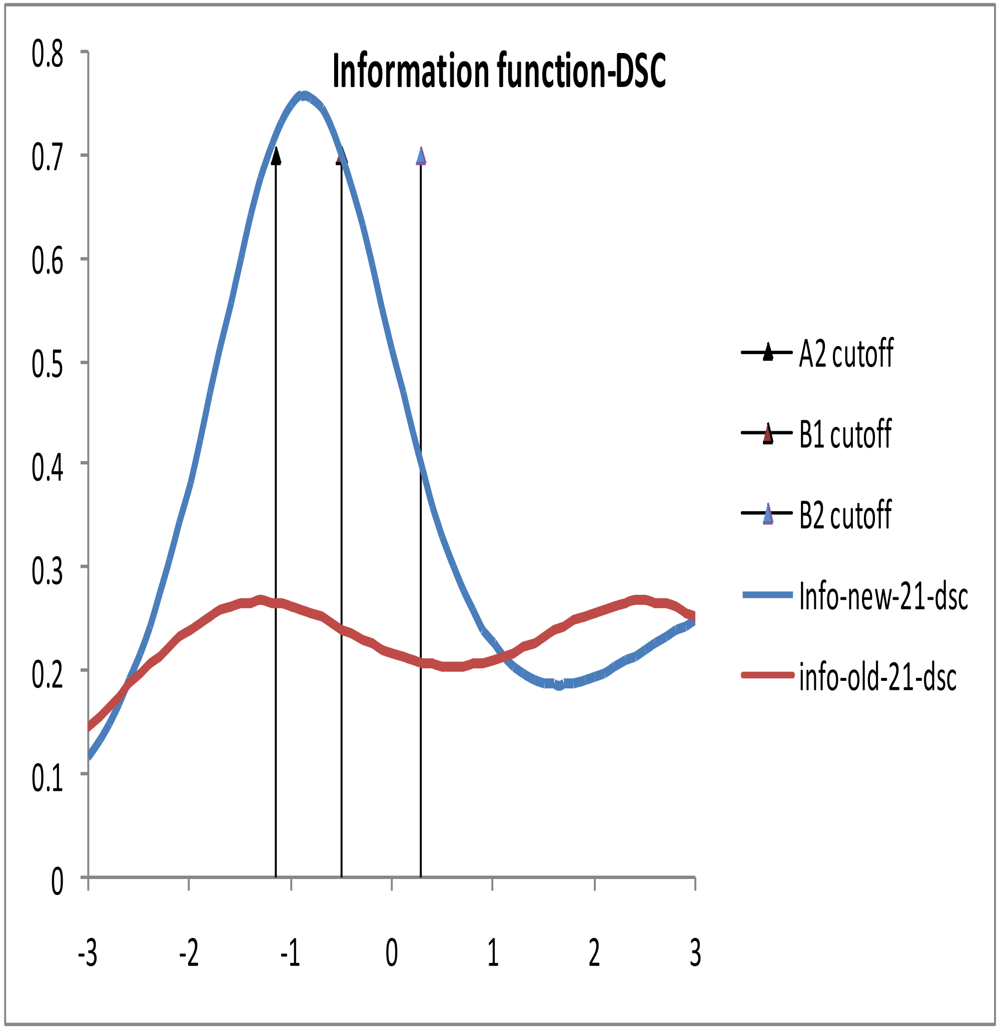
Results

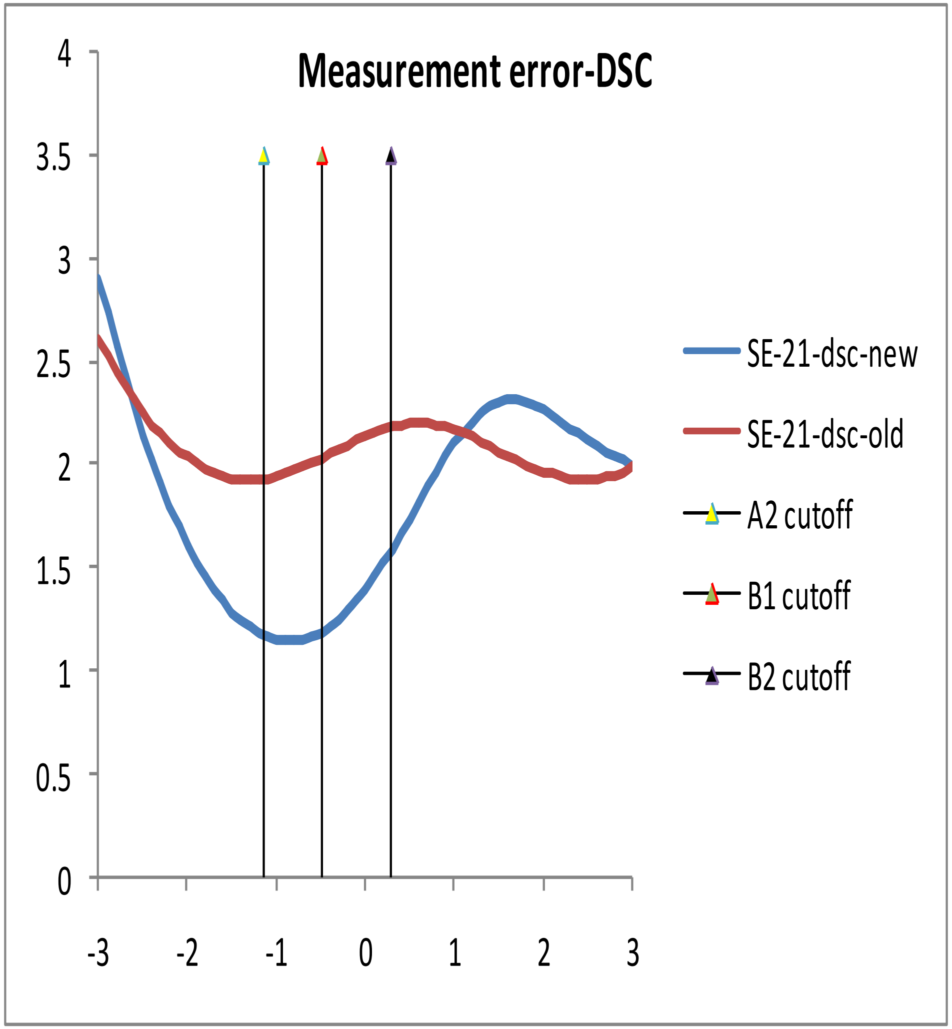
Two separate IRT analyses were performed comparing the effects of current and new transformation methods

The ***information function*** much higher in the most relevant areas (CEF A2 to B2).

➡ The ***conditional error of measurement*** reduced substantially in this most relevant area

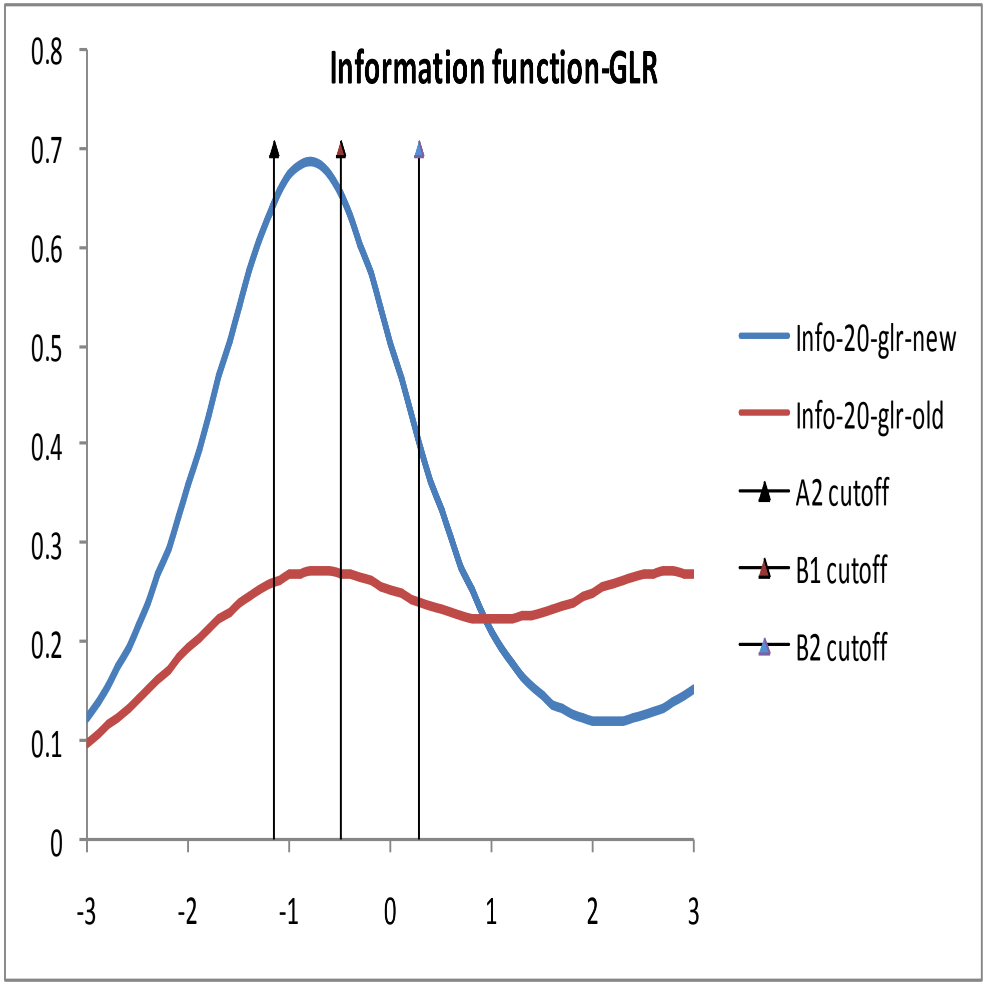
Trait-level: *Development, Structure & Coherence*

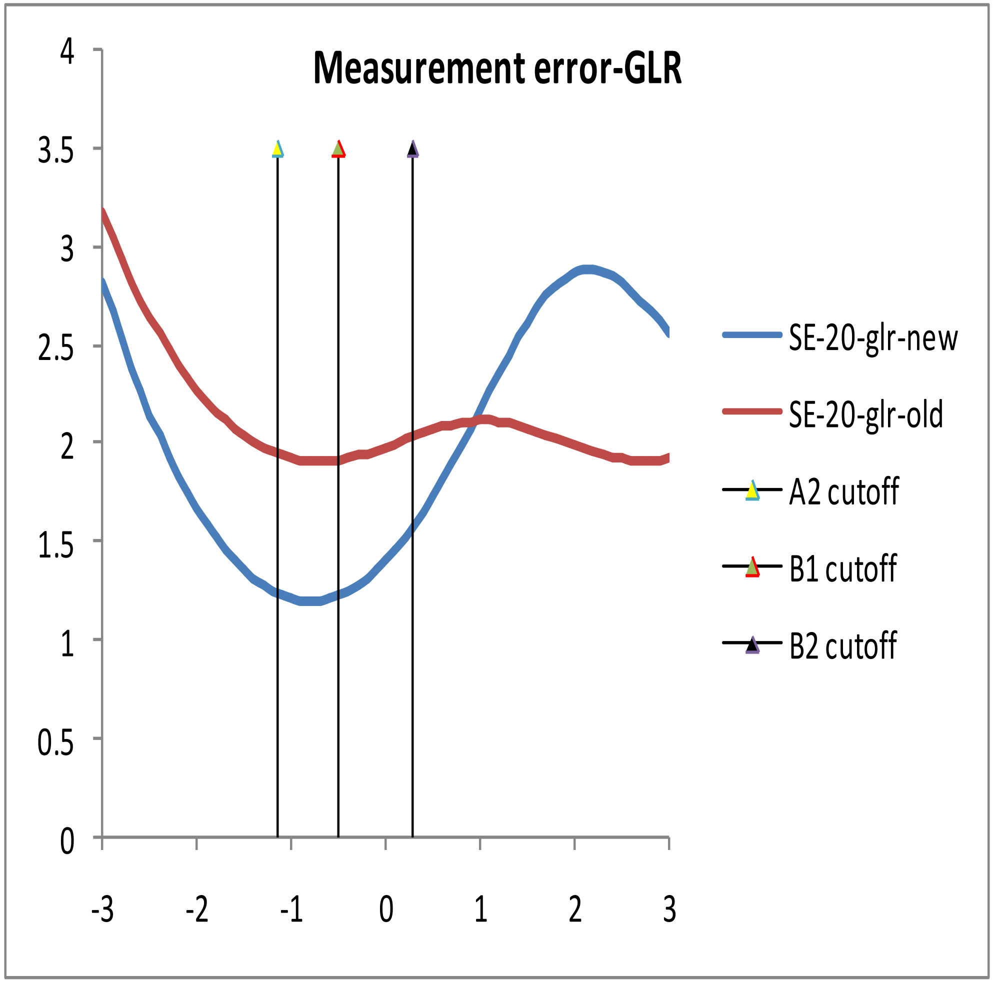




PEARSON

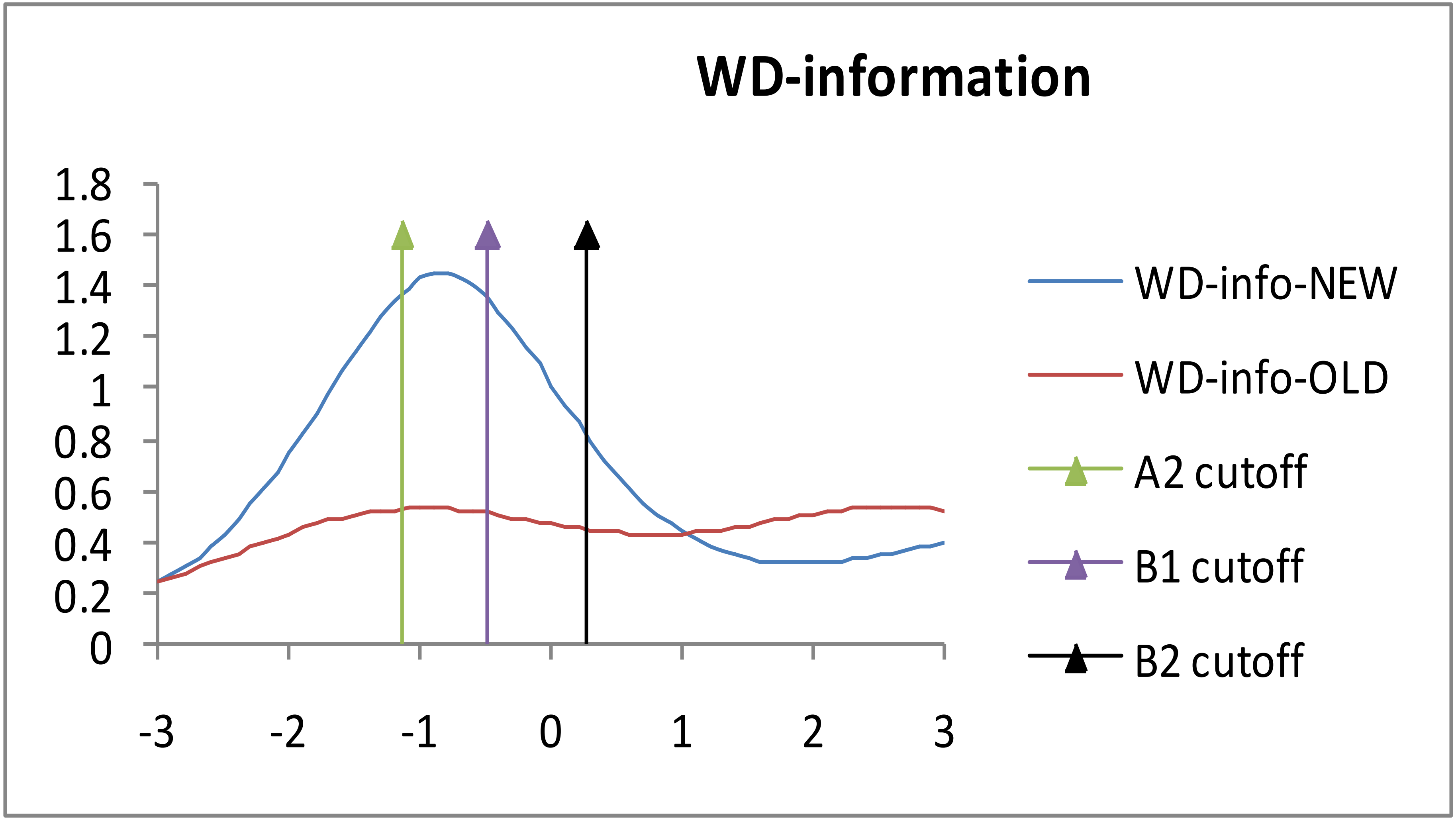
Trait-level: *General Linguistic Range*



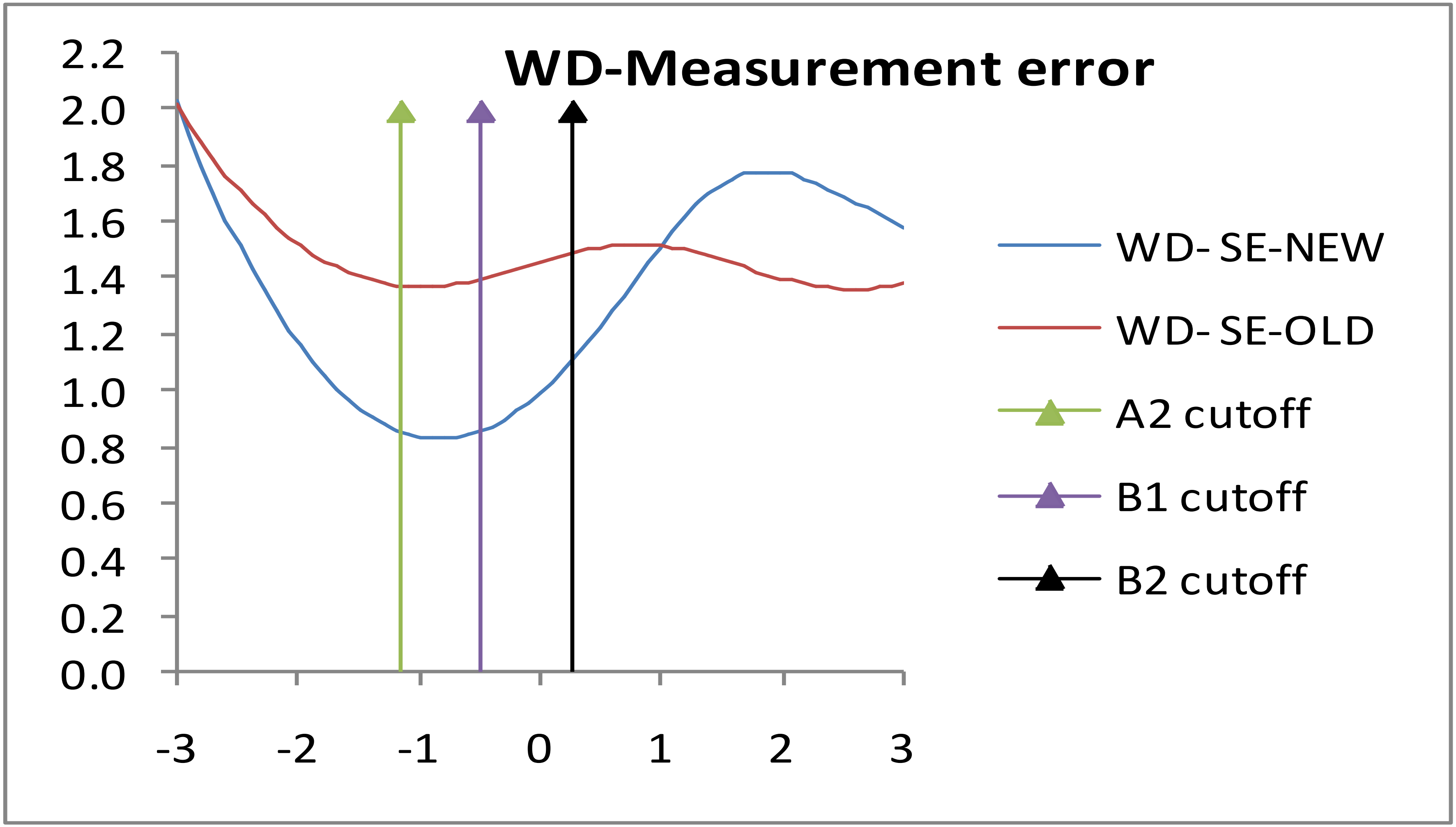


PEARSON

Item-level: *Written Discourse*



Item level: *Written Discourse*



CEM on *Written Discourse* Scaled Scores

	theta	scaled score (new)	theta	scaled score (old)
A2	-1.68	21.85	-2.37	8.93
	0.02	53.60	0.38	60.37
CEM		**15.87**		**25.72**
B1	-0.71	39.97	-0.74	39.34
	1.01	72.09	2.05	91.60
CEM		**16.06**		**26.13**
B2	1.04	72.67	0.38	60.44
	3.29	114.75	3.37	116.29
CEM		**21.04**		**27.93**

Analysis Phase II: Improving *Grammar* and *Vocabulary*

Background

During Field Tests, trait scores of ***'Intonation***' & '***Language Use***' for certain item types were collected.

Proposed two-step approach:

Step 1

Only change current machine estimates transformation

Step 2

Change current machine estimates transformation **AND** add new traits

Proposed New Traits

Item types	Trait scores	Enabling skills
07-SR-READ	Intonation	Grammar
19-SS-DESC	Language use	Grammar
20-LS-PRES	Language use	Grammar
16-LS-REPT	Content	Grammar
19-SS-DESC	Content	Vocabulary
20-LS-PRES	Content	Vocabulary

Three IRT analyses were carried out:

Base analysis

This analysis used the original trait scores on a 3-point scale

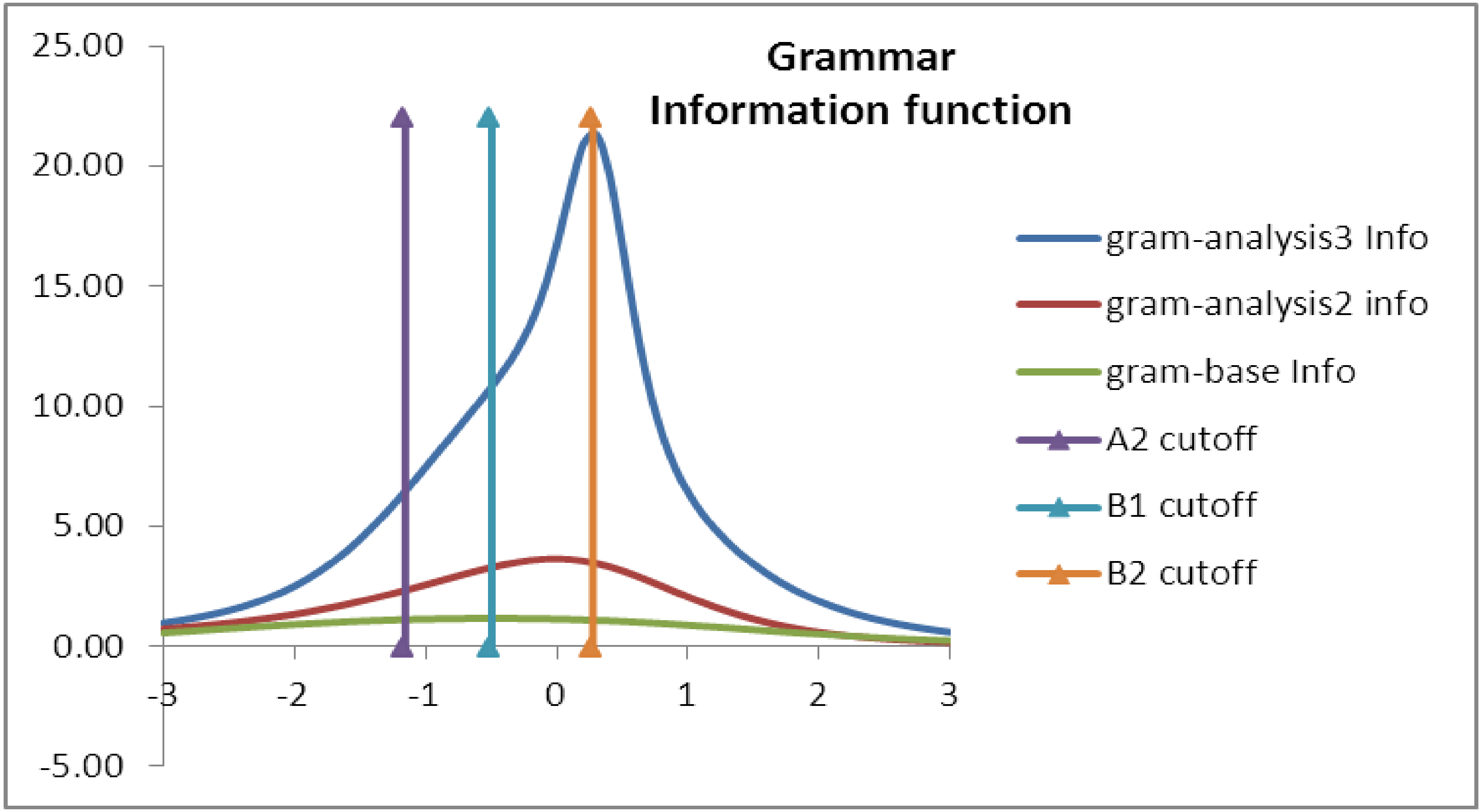
Analysis 2

The second analysis used the original trait scores but on a 4-point scale

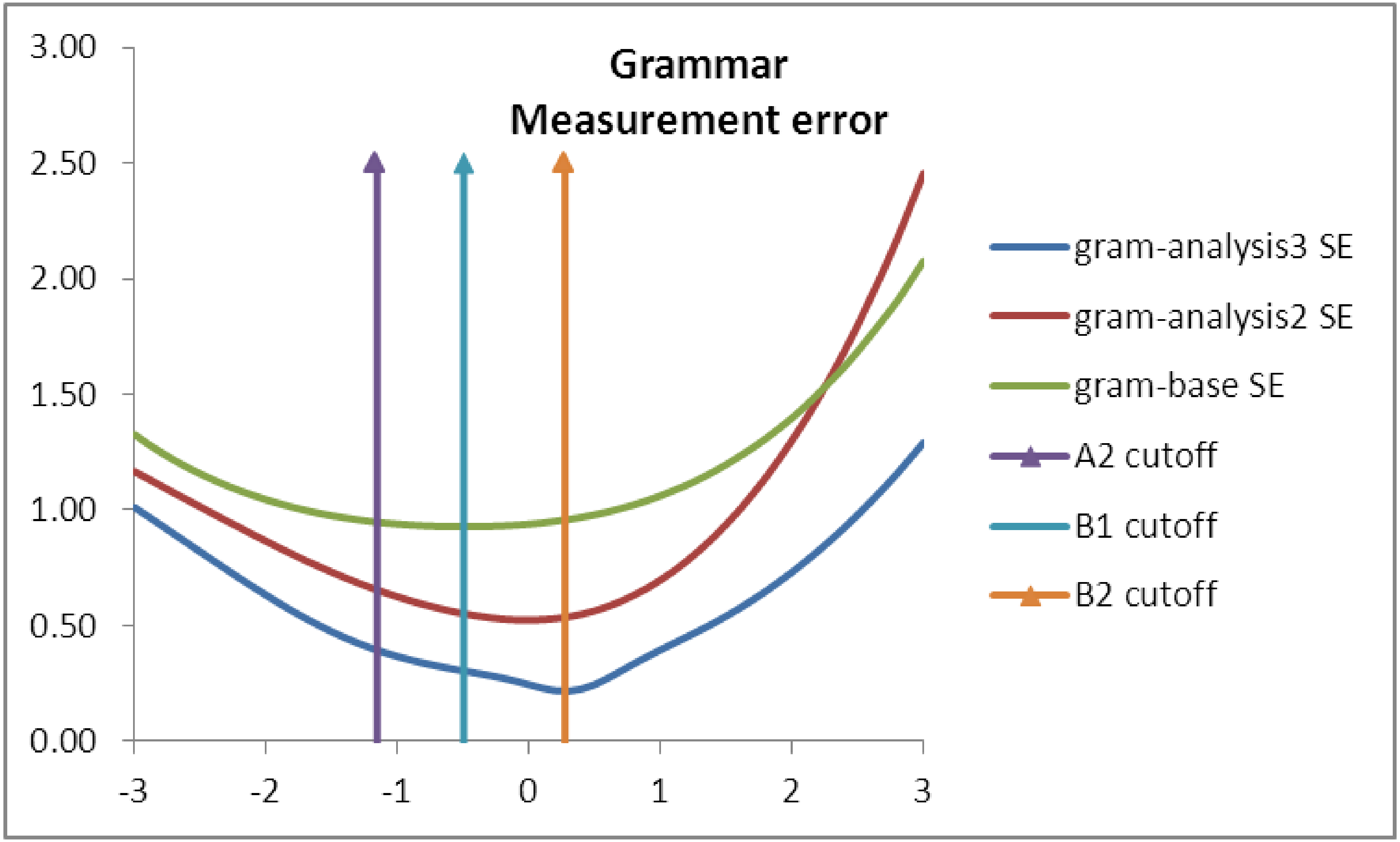
Analysis 3

The third analysis used a 4-point scale and also included 6 new traits, four for ***Grammar*** and two for ***Vocabulary***

Grammar Information function comparison

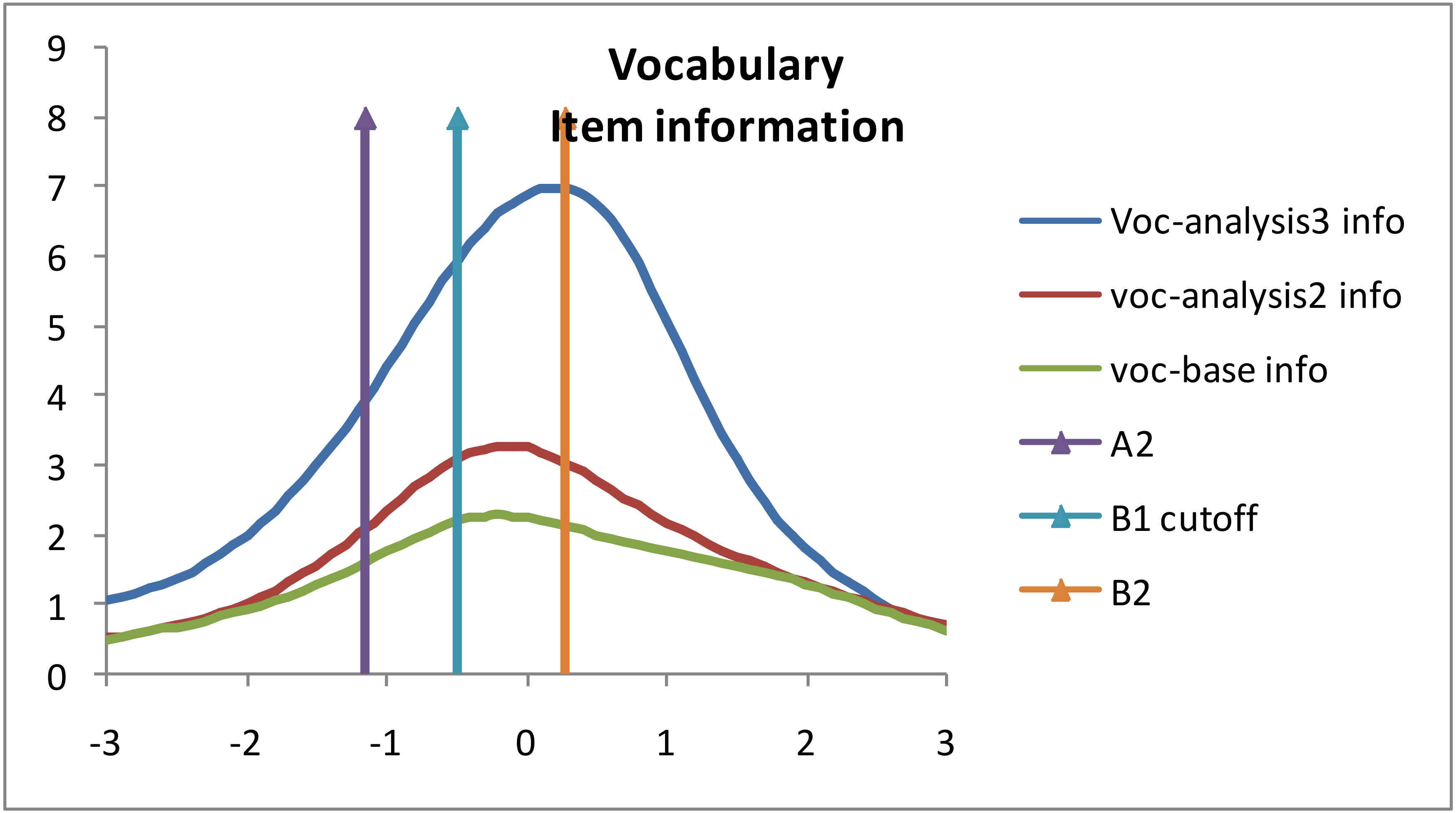


Grammar Measurement error comparison

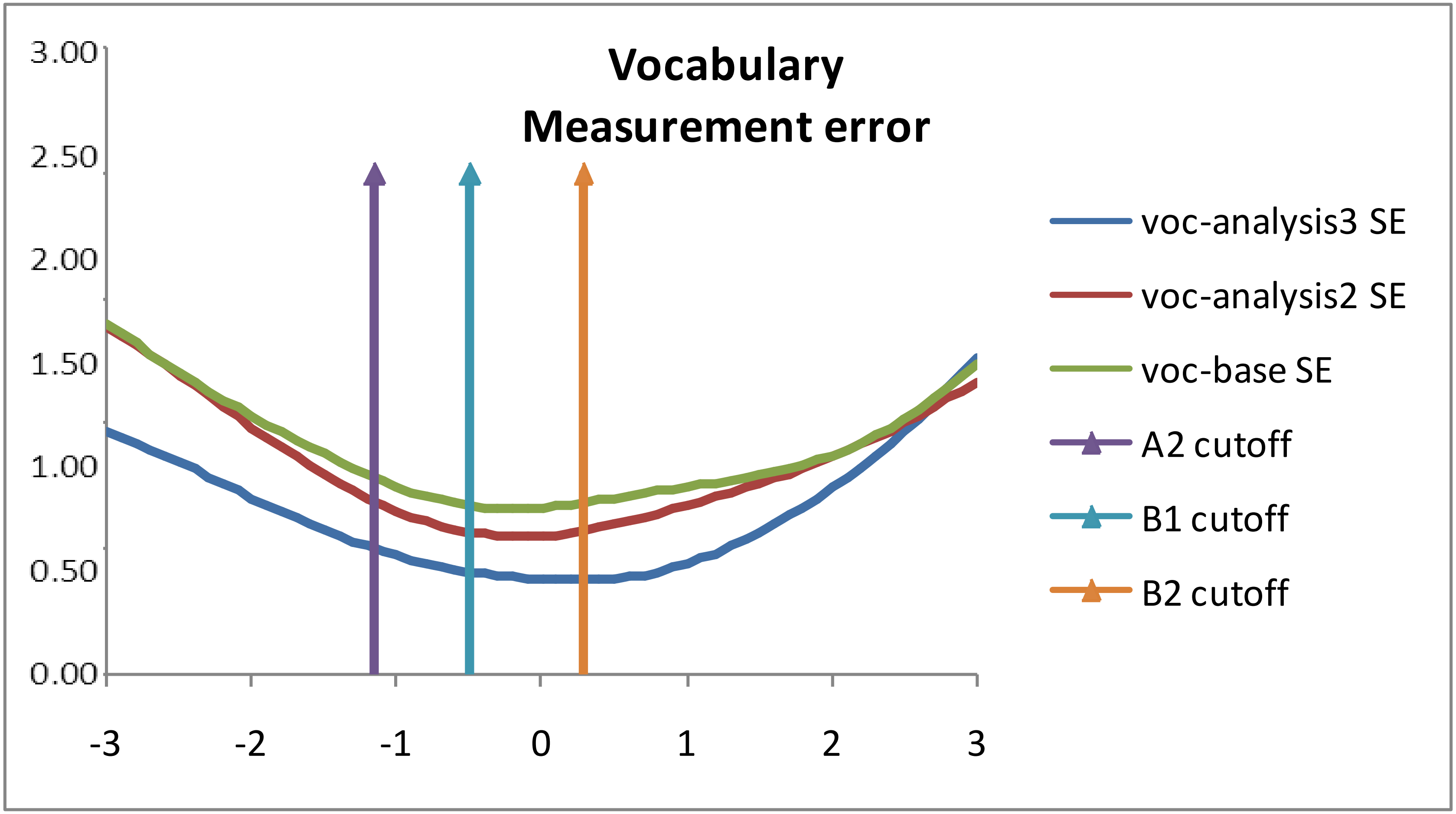


PEARSON

Vocabulary Item Information comparison



Vocabulary measurement error comparison



CEM on Vocabulary and Grammar scores

	Base analysis		Analysis2		Analysis 3	
	Voc.	Gram.	Voc.	Gram.	Voc.	Gram.
A2	14.49	17.67	12.67	12.04	9.23	7.17
B1	12.63	17.35	10.65	10.33	7.67	5.69
B2	12.86	17.95	10.76	10.07	7.08	4.06

Conclusion

Two ways to improve measurement precision:

1. Break the scores into more score categories;
2. Include more traits

As the same scores which contribute to the enabling skills scores are also used to compute the item scores, which in turn are used to generate the overall score and the four communicative skill scores (Listening, Speaking, Reading, and Writing), reducing the error on the enabling skills scores will also reduce the measurement error on the overall and the communicative skill scores.

Concluding remarks

In the documentation on the test it is explicitly stated that decisions on university admission should **only** be based on the overall score and the four communicative skill scores.

The enabling skills scores are provided to inform test takers and their teachers on possible causes why students are not obtaining the overall and communicative skill scores they need to be admitted to the institution of their choice.

Increasing the accuracy of the enabling skill scores would make this information more valuable and test takers would be able to direct their learning efforts more efficiently at improving their English language competencies.

Final remark

Evidence is available that machine scores have higher correlation with better trained human raters.

Therefore the following statement is warranted:

> Once they are validated, machine scores are superior to human scores and human scores are no longer the model to be targeted by machine scores.