Improving the BECC Bunkyo English Tests **Searching for Validity**



A three-stage process of analysis combining Rasch Analysis, Excel Analysis, and Text Analysis.

"Validity is the most fundamental consideration in developing and evaluating tests" (APA, AERA, NCME, 2014) "The key criterion driving assessment at Cambridge Assessment is validity"

Why?

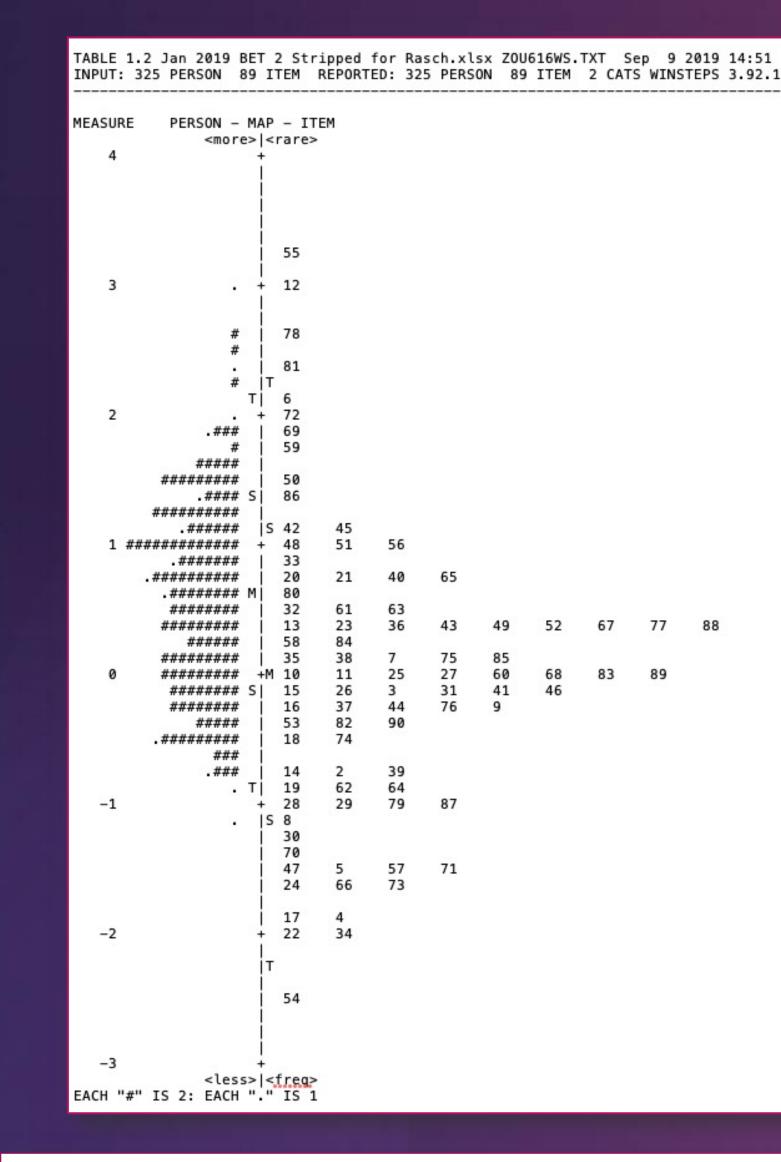
(Shaw, 2020)

1. Rasch Analysis

2. Excel Analysis

3.Text Inspector

- Person / Item Variables
- Item Correlation
- **Item Statistics**



- The Excel BET Database
- **Excel Question Analysis**
- **Excel Microanalysis**
- **Excel Testlet Analysis**

Que	stions		To	tals		8		A1-A	2 Stream Cl	asses					A2-B1 Stre	am Classes		
Section	Question	Total Average	A1-A2 Stream Avg	A2-B1 Stream Avg	High Stream	FE1 (A1-A2)	FE3 (A1-A2)	FE6 (A1-A2)	FE7 (A1-A2)	FE9 (A1-A2)	FE11 (A1-A2)	FE12 (A1-A2)	FE2 (A2-B1)	FE4 (A2-B1)	FE5 (A2-B1)	FE8 (A2-B1)	FE10 (A2-B1)	FE13 (A2-B1)
BERT 1	2	49%	30%	64%	A2-B1	39%	33%	30%	19%	43%	25%	19%	53%	83%	73%	70%	67%	47%
BERT 1	3	72%	57%	84%	A2-B1	61%	67%	59%	46%	46%	50%	67%	80%	83%	97%	87%	93%	97%
BERT 1	4	82%	73%	88%	A2-B1	71%	81%	67%	77%	79%	64%	70%	90%	83%	93%	83%	90%	97%
BERT 1	5	82%	68%	94%	A2-81	50%	70%	74%	62%	86%	64%	67%	87%	100%	100%	97%	100%	97%
BERT 1	6	36%	23%	46%	AZ-B1	25%	30%	15%	19%	32%	25%	15%	43%	63%	63%	57%	43%	30%
BERT 2	7	42%	43%	41%	A1-A2	29%	63%	63%	42%	39%	21%	48%	43%	50%	57%	33%	33%	40%
BERT 2	8	55%	38%	68%	A2-81	36%	33%	41%	46%	46%	29%	37%	50%	63%	77%	83%	77%	63%
BERT 2	9	74%	57%	88%	A2-B1	43%	74%	63%	54%	54%	46%	63%	87%	87%	83%	97%	93%	90%
BERT 2	10	56%	39%	69%	A2-B1	39%	41%	22%	46%	29%	43%	52%	47%	63%	83%	80%	83%	67%
BERT 2	11	48%	18%	71%	A2-B1	25%	19%	26%	15%	18%	14%	11%	60%	83%	77%	80%	83%	63%
BERT 3	12	11%	10%	11%	A2-B1	11%	7%	15%	8%	11%	0%	19%	7%	7%	7%	13%	13%	13%
BERT 3	13	42%	32%	50%	A2-81	29%	30%	22%	38%	18%	46%	41%	20%	57%	37%	53%	60%	53%
BERT 3	14	36%	26%	45%	A2-B1	32%	33%	11%	15%	39%	29%	19%	33%	50%	43%	60%	60%	37%
BERT 3	15	77%	66%	86%	A2-B1	54%	89%	44%	69%	71%	68%	67%	83%	90%	97%	77%	97%	90%
BERT 3	16	71%	50%	88%	A2-B1	39%	33%	52%	54%	57%	54%	59%	83%	100%	93%	90%	93%	77%

Test	Info					BET 1 (Apr	il 2019)		
			Low (A1-A2)	High (A2-B1)				Outlier? [Low stream higher or	Ability Determiner? [High stream highest
Section -	Question T	Overall .	Stream %	- Stream % -	Easiest?	- Hardest?	- (High - Low) -	close to High stream] 🝷	over Low stream] 💌
BERT Part 1	2	49%	30%	64%			34%		T-6th biggest AD
BERT Part 1	3	72%	57%	84%			28%		
BERT Part 1	4	82%	73%	88%	5th easiest		16%		
BERT Part 1	5	82%	68%	94%	4th easiest		27%		
BERT Part 1	6	36%	23%	46%			23%		
BERT Part 2		42%	43%	41%			-2%	4th biggest outlier	
BERT Part 2	8	55%	38%	68%			30%		
BERT Part 2	9	74%	57%	88%			32%		T-8th biggest AD
BERT Part 2	10	56%	39%	69%			30%		
BERT Part 2	11	48%	18%	71%			53%		1st biggest AD
BERT Part 3	12	11%) 10%	11%		(1st hardest	1%	T-5th biggest outlier)
BERT Part 3	13	42%	32%	50%			18%		
		22.07		112222			10.000		

- **BET Text Analysis**
- **Reading Section Analysis**
- Listening Section Analysis

The GEAC committee wants to create assessments that become progressively challenging through each section of the test, therefore the CEFR level progresses from A2 to B1+.

This is the overall Lexical P given as a percentage, with Academic text.				
For the meaning of A1-C2 p	lease see here.			
PERCENTAGE	NUMBER OF METRICS USED	CEF	RLEVE	L
30.95	14		B1	
Scorecard				
Scorecard These are the metrics used Listening text. These metric statistically significant in d If more metrics are used the	is have been identified istinguishing between	d in our rese n different L	earch as Istening	
These are the metrics used Listening text. These metric statistically significant in d	is have been identified istinguishing between	d in our rese n different L	earch as Istening	levels.
These are the metrics used Listening text. These metric statistically significant in d If more metrics are used the	ts have been identifie istinguishing betwee e score is likely to be	d in our rese n different L	tarch as Istening In. O	levels.
These are the metrics used Listening text. These metric statistically significant in d If more metrics are used th Statistics	ts have been identifie listinguishing betwee e score is likely to be PER SENTENCE	d in our rese n different L more reliabl	open /	levels.
These are the metrics used Listening text. These metric statistically significant in d If more metrics are used the Statistics AVERAGE SYLLABLES	e score is likely to be PER SENTENCE	d in our rese n different L more reliabl 11.96	open /	levels.
These are the metrics used Listening text. These metric statistically significant in d If more metrics are used the Statistics AVERAGE SYLLABLES AVERAGE SYLLABLES	e score is likely to be PER SENTENCE	d in our rese n different L more reliabl 11.96 1.29	OPEN A2 81	levels.

	TOTAL	TOTAL		MODEL IN	FIT OUT	FIT PTMEAS	UR-AL EXACT	MATCH
	SCORE	COUNT	MEASURE			ZSTD CORR.		
2	249	325	75	.14 .92	-1.2 .84	-1.6 .40		
3	210	325	11	.12 .88	-2.6 .83	-2.71 .47	.32 72.0	67.9
4	293	325	-1.85	.19 1.01	.1 .98	.0 .19	.20 90.2	90.1
5	284	325	-1.56	.17 .91	71.74	-1.6 .36	.22 87.4	87.4
6	65	325	2.12	.15 .96	4 .97	21 .33	.29 81.8	80.5
7	195	325	.11	.12 1.02	.4 1.02	.41 .30	.33 65.5	66.0
8 9	268	325	-1.14	.15 .96	5 .96	31 .30	.25 82.5	82.5
9	216	325	20	.12 1.12	2.5 1.15	2.1 .15	.31 64.6	68.9
10	203	325	.00	.12 .88	-2.8 .85	-2.71 .48	.32 73.2	67.0
11	204	325	02	.12 1.08	1.9 1.12	1.9 .21	.32 63.1	67.1
12	34	325	2.94	.19 1.18	1.3 1.80	3.5110	.23 89.2	89.5
13	172	325	.44	.12 .90	-2.7 .88	-2.71 .46	.34 70.2	64.5
14	248	325	73	.14 .97	5 .97	31 .32	.28 78.5	76.4
15	211	325	12	.12 .88	-2.8 .81	-3.11 .49	.32 71.1	68.1
16	219	325	25	.12 1.06	1.2 1.07	1.01 .24	.31 67.4	69.4
17	294	325	-1.89	.19 1.06	.5 1.19	.91 .09	.19 90.5	90.5
18	238	325	56	.13 .94	-1.1 .89	-1.2 .38	.29 75.7	73.7

ITEM	DATA CODE	SCORE VALUE	DATA COUNT %	AB MEAN	ILITY P.	SD	S.E. MEAN	INFT MNSQ		PTMA CORR
12	A B C		7 2 284 87 34 10	.03 .61 .37		.76	.25 .05 .12			11 .13 10
	12	What do you t	hink number 1 is?	A B C	Thank y I think ti Let's as	hat's r	-	e.		
12	What	do you thi	nk number 1 is?		AB			u for te at's rig	-	ne.

BERT Part 3	16	71%	50%	88%		38%	4th biggest AD
BERT Part 3	15	77%	66%	86%	8th easiest	20%	
BERT Part 3	14	36%	26%	45%		19%	

	Question	Comment
1	Sayaka met a police officer, and <u>25</u> police officer asked Sayaka to describe her earrings. She said they were gold and shiny. The officer then asked <u>26</u> contact information, <u>27</u> she gave <u>28</u> her name and her address.	Difficult, but accurately determining ability
2	26 A for Sayaka to B Sayaka for her C to Sayaka 15 How's the weather there today? A I'll go cycling. B I am hot. C It is foggy.	Easy, but accurately determining ability
3	11 Kumiko wants to enough money to travel to America. A earn B cost C give	Very highly determining ability
4	7 She has worked three days for her new A employer B employee C employment	Very weakly determining ability (or undermining it)

Section	Section	Overall	Difficulty Ranking	Low Stream	High Stream	High - Low	Ability Determine Ranking
	1	65.6%	14	53.4%	73.1%	19.7%	7
	2	65.5%	13	51.9%	74.3%	22.4%	3
	3	56.6%	7	43.6%	64.2%	20.6%	5
Reading	4	62.9%	10	46.6%	72.6%	26.0%	2
(BERT)	5	64.8%	11	51.9%	72.9%	21.0%	4
	6	54.7%	4	43.7%	61.7%	17.9%	8
3	7	43.7%	1	26.6%	54.3%	27.7%	1
	8	58.2%	8	48.4%	64.9%	16.4%	11
	1	56.1%	6	45.7%	61.5%	15.8%	12
	2	53.6%	2	43.1%	61.0%	17.9%	9
Listening	3	65.2%	12	52.4%	73.0%	20.5%	6
(BELT)	4	55.0%	5	46.0%	60.4%	14.4%	13
	5	54.7%	3	43.9%	61.0%	17.1%	10
	6	58.3%	9	51.9%	62.9%	11.0%	14

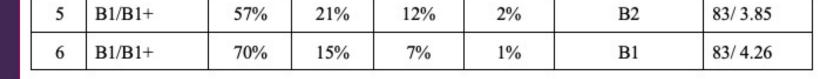


		BET 11	Reading Voc				
Part	CEFR Aim	A1	A2	B1	B2	Overall CEFR	FRE/FKL
1	A2	73%	27%	-	-	NA	NA
2	A2/ A2+	61%	18%	15%	121	NA	NA
3	A2/ A2+	69%	7%	5%	121	NA	90 / 1.8
4	A2/ A2+	80%	7%	3%	-	A1+	89 / 2.3
5	A2/B1	66%	22%	5.5%	1%	A2	72 / 6.5
6	A2/ A2+	57%	25%	11%	-	A2+	80 / 5.5
7	B1/B1+	63%	17.5%	9%		A2	82 / 4.4
8	B1/B1+	60%	14%	7%	-	B1	65 / 8.5

		BET	2 Listening	Vocabulary			
Part	CEFR Aim	A1	A2	B1	B2	Overall CEFR	FRE/FKL
1	B1	58%	17%	8%	1.5%	B1	83/4.0
2	A2	66%	14%	3.5%	1%	B1+	85/ 3.25
3	A2	73%	10%	1%		A2+	86/3.0
4	A2	62%	14%	6%	5%	B2	83/ 3.85
	2	8			2		1

Section	Section	Overall	Difficulty Ranking	Low Stream	High Stream	High - Low	Ability Determiner Ranking
	1	65.6%	14	53.4%	73.1%	19.7%	7
	2	65.5%	13	51.9%	74.3%	22.4%	3
	3	56.6%	7	43.6%	64.2%	20.6%	5
Reading	4	62.9%	10	46.6%	72.6%	26.0%	2
(BERT)	5	64.8%	11	51.9%	72.9%	21.0%	4
	6	54.7%	4	43.7%	61.7%	17.9%	8
	7	43.7%	1	26.6%	54.3%	27.7%	1
	8	58.2%	8	48.4%	64.9%	16.4%	11
	1	56.1%	6	45.7%	61.5%	15.8%	12
	2	53.6%	2	43.1%	61.0%	17.9%	9
Listening	3	65.2%	12	52.4%	73.0%	20.5%	6
(BELT)	4	55.0%	5	46.0%	60.4%	14.4%	13
	5	54.7%	3	43.9%	61.0%	17.1%	10
	6	58.3%	9	51.9%	62.9%	11.0%	14

Conclusions



"No test is valid for all purposes or in all situations or for all groups of individuals." (Shaw 2020)

All we can do is try to to provide evidence of having performed due process to satisfy the following stages of construct validity:

- The contents of the tests match the contents of the curriculum
- Students answer the questions in the manner intended, and their answers to different questions relate in an expected way ightarrow
- Users can interpret results in the manner intended ightarrow



Jordan Svien Hiroshima Bunkyo University jsvien@h-bunkyo.ac.jp

Tyler Montgomery

Kyushu Sangyo University p-421492@mail.kyusan-u.ac.jp

