広島文教大学紀要　55，2020

【原著】

Improving the BECC Bunkyo English Tests in the Search for Validity

Richard Sugg, Jordan Svien and Tyler A. Montgomery

文教大学 BECC 英語テストの妥当性の向上

Richard Sugg, Jordan Svien and Tyler A. Montgomery

Abstract

Developing in-house streaming and yearly assessment tests is a constantly evolving process for most educational institutions. Six years ago, in line with our decision to base our General English (GE) curriculum on the CEFR, teachers at Hiroshima Bunkyo University's Bunkyo English Communication Center (BECC) developed a new streaming test for incoming students. It was hoped that this Bunkyo English Test (BET) could be proven to be a valid test of our students' abilities, and also be used to track student CEFR reading and listening levels and progress over two years. This report will briefly outline the test's original development background, before going over the three-stage Rasch, Excel and Text Inspector analysis process that has evolved to form the basis of its yearly review and rewriting. Examples of the results of the process will be given, along with ideas on how we can progress with the next stage of our tests' development.

Introduction

Six years ago, the BECC started aligning its General English (GE) curriculum and assessments with the CEFR over what was at that time a compulsory first two years of English study at the university. The Bunkyo English Tests (BETs) that arose from this decision are "institutional standardized reading and listening tests administered as part of the GE curriculum" (Bower, J. et al, 2014), and were designed to stream students (BET 1), and to track their progress (BETs 2 and 3). They also had a goal of being able to give the students some form of CEFR certification indicating their level of performance at the end of the two years (BET 3). (For a greater understanding of the BETs and their creation, see Bower, J. et al, 2014).

Creating the BETs is an ever-evolving and iterative process. As members of the General English Assessment Committee (GEAC), we analyze the results of previous tests and identify questions that need to be revised, rewritten, or removed completely from the next version of the test. In 2014 and 2015, the BETs were only analyzed using Rasch Analysis. Also, in 2015, the original BET format of five reading parts, two vocabulary and grammar questions and five listening questions outlined by Bower, J. et al, undertook its first minor change, in that the two vocabulary and

grammar questions were incorporated into the reading section (BERT). This raised the number of test items from the original 66 items to 68. In 2016, to widen the spread of results generated by Rasch Analysis, and with an eye on future cut score setting for CEFR ability ranking, one more reading question and one more listening question were added, increasing the number of sections to eight and the total number of items to 86. Finally, in 2017, the number of items in Part 1 of the listening section (BELT) was increased to ten, bringing the total to 89 items. The time allowed was also increased from an original 60 minutes in 2014 to 75 minutes in 2017. From 2016, the BET results were also added to an Excel Analysis Database, and during the 2019 round of analysis, we also added text analysis via the online Text Inspector site. It is this new three-stage process that we will outline in the rest of this paper.

A Word on Validity

In all testing situations, test writers strive to make their tests 'valid'. Indeed, the renowned Standards for Educational and Psychological Testing states that "Validity is the most fundamental consideration in developing and evaluating tests" (APA, AERA, NCME, 2014). Also, Cambridge Assessment, whose suite of English certification exams are widely used around the world, and from which the KET and PET tests we use as a guide when designing our assessments and rubrics are taken, says that "The key criterion driving assessment at Cambridge Assessment is validity" (Shaw, 2020). What is not so readily agreed upon is what exactly test validity is. Are we talking about individual questions being valid, the overall test being valid, the test's score being valid, or how the test score is used being valid? Shaw (2020) divides those who try to define validity into three camps:

1) The Conservatives, who, in quoting Borsboom, Mellenbergh and Van Heerden (2004), Shaw (2020) says believe that "A test is valid if it measures what it purports to measure."
2) The Liberals, who he says believe that "Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (Messick, 1989)"
3) The Traditionalists / Moderates, who, in quoting the Standards (APA, AERA, NCME, 2014), he says support the idea that "Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests."

In the GEAC it could be argued that we currently fall somewhere between the Conservatives and the Liberals when we use the term validity. This is because like the conservatives, we use it to mean that our tests measure what they are supposed to measure: in BETs 2 and 3, a student's ability to answer questions based on the materials they have studied. However, like the Liberals, we also want to a) show validity when we talk about the actions we take based on the test scores: deciding from BETs 1 and 2 what stream a student should go in, and also b) when we talk about the interpretations we draw from our test scores: where a student is in terms of their CEFR level. How then, do we validate our tests? With over 90 different forms of specialized validity, it is very

difficult to know where to start. Previously we have concentrated on criterion validity and content validity. 'Criterion validity' looks at the correlation between our BETs and other similar, already well-proven English tests. In our case, we have modelled our questions on the Cambridge Assessment KET and PET tests, which we know cover the CEFR levels from A1 up to B1. We can then use student results to divide our students into two streams: A1–A2 and A2–B1.

This in turn brings us to 'content validity', which was used as the basis for writing the original BETs. Bower, J. et al (2014) states that for a test to have validity, "it must cover a broad sample of content from its target domain (Kane, 2013). This is traditionally referred to as 'content validity'. For the BETs, the target domain is the lesson handouts for the GE course."

This is a very common and sensible approach, but even this has its issues. Shaw (2020) asks, "What if your questions fail to tap the intended proficiency? What if the intended proficiency was tapped, and demonstrated, but not rewarded appropriately? What if test behaviour fails to generalize?"

In an attempt to unify and improve on both of the above definitions of validity, those involved in educational assessment are increasingly using the term 'construct validity' to describe what a test should do. Standards (APA, AERA, NCME, 2014) defines five sources of evidence that must be satisfied: the test content, the response process, the internal structure, the relations to the variables and the consequences of testing. This requires asking ourselves if the content of the tests match the curriculum content, if the students answer the questions in the manner intended, if the tests are marked in the way they are intended to be, if the students' answers to different questions relate in a way we would expect them to, and whether or not users of the results can interpret them in the manner intended.

Rasch Analysis

Rasch analysis is a worldwide recognized method of analyzing students' results which anyone can access and learn about via the Institute for Objective Management, Inc, and their website at https://www.rasch.org/. Rasch analysis "provides a mathematical framework against which test developers can compare their data. The model is based on the idea that useful measurement involves the examination of only one human attribute at a time on a hierarchical 'more than/less than' line of enquiry." (Bond and Fox, 2001)

To run Rasch via the Winsteps (Linacre, 2008) program, the BETs answer keys and the raw BET results are prepared in Excel files. These are then used to create a Winsteps control file. When this file is dropped into Winsteps, the program then first calculates the fit statistics (Figure 1). First, we check the input to make sure that the numbers of students tested and of questions answered (items) are correct. After this we look at 3 elements:

- Person separation: This indicates the range of person ability. Anything over two is good and is what we need to decide on two streams of classes. In the example below the range

of 2.71 is adequate for streaming purposes.

- Item separation: This indicates the range of item difficulty. Again, anything over two is good, so in our example below taken from BET 2 taken in January 2019, a range of 7.96 is pleasing to see.
- Reliability: This will always be higher for items than for people in our case because each item is measured by every student, while each student is only measured by the number of items.

Figure 1. *BET 2 2019 Fit Statistics*

```
INPUT: 325 PERSON  89 ITEM  REPORTED: 325 PERSON  89 ITEM  2 CATS WINSTEPS 3.92.1
-------------------------------------------------------------------------------
PERSON: REAL SEP.: 2.71  REL.: .88 ... ITEM: REAL SEP.: 7.96  REL.: .98
```

The program then provides 36 different output tables. For our analyses, we concentrate on just three:

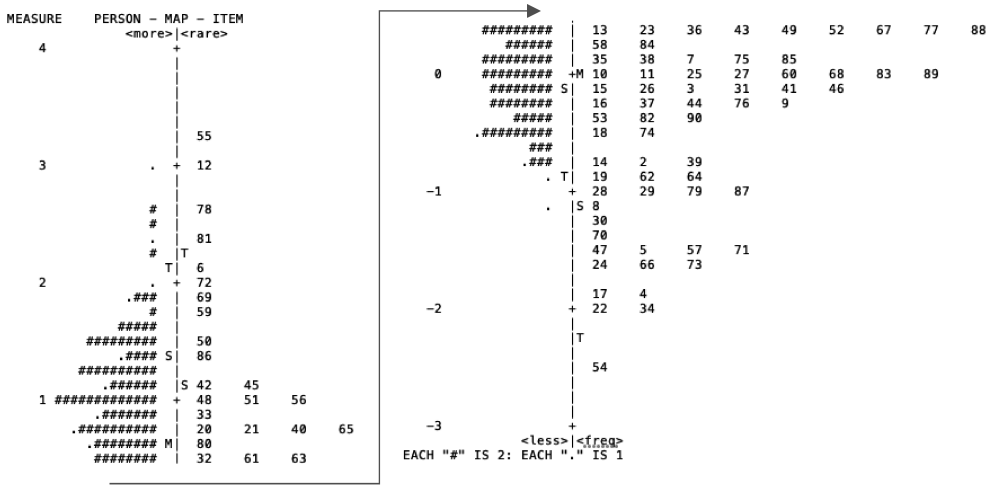
- Output Table 1: Person / Item maps, which show how many students are getting what questions correct or incorrect.
- Output Table 26: Item Correlation, which shows us any problem items that have low point measure correlations.
- Output Table 14; the Item Statistics, which show how individual items are performing.

By combining the results of all three, we can identify individual test items that need to be re-written or removed entirely from the tests.

Person / Item Variable Maps

Output Table 1 contains 13 variable maps, of which we find map 1.2 (Figure 2) to be the most useful.

Figure 2. *BET 2 2019 Output Table 1 Variable Map 2 (cropped for publication)*



Students on the left are represented by hashes or dots, with a hash mark equaling two students and a dot one. Test items are on the right, while the scale in the middle is shown in Rasch logits. The more logits, the higher the student ability and the difficulty of the question. However, it is important to note that students are placed where the model predicts they have a 50% chance of getting an item correct. The map does not tell us whether test-takers got questions at the same logit value correct, only their probability of doing so. A statistically 'ideal' test would have two perfectly uniform columns of people and items. Although this is never actually possible, what the map clearly shows is:

- Whether there are any students at the top of the chart who are not being adequately tested. This is known as a 'ceiling effect'.
- If there are too many or too few items anywhere.
- If there are too many easy items not measuring anything.

In Figure 2, we can see that while no students are outperforming the test, there is a group of eight students who are coming close to doing so. This would indicate that either these students have an ability of CEFR B1 or above already, or that we have not written enough 'difficult' questions. Also, there are 14 items at the bottom of the map that are effectively not testing anyone. A few easy questions at the beginning of the reading and listening sections are acceptable, as we do also have to assess just how low some very low students are, but this is too many. Overall, in this test we seem to have too many students clumped together with too many similar level items. As the BET 2 measures only the work done during year 1 (whereas BETs 1 and 3 are a mixture of items from both years 1 and 2), this can be expected, but also indicates that we need to rewrite or remove certain items to increase the spread of student separation.

Item Correlation

The item correlation tables list the items in order of point measure correlation, with the worst-performing item at the top. Output table 26.1 (Figure 3) indicates that question (Item) 12 has serious issues (discussed later in this paper), while the questions listed after that need to be analyzed more closely. Any of these questions that also correspond with the outlying questions in output table 1.2 are the first to be checked in greater depth.

Figure 3. *BET 2 2019 Output Table 26.1 Item Correlation (cropped for publication)*

ITEM	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%
12	34	325	2.94	.19	1.18	1.3	1.80	3.5	-.10	.23	89.2	89.5
45	122	325	1.14	.12	1.22	4.5	1.27	4.4	.05	.33	57.8	67.2
48	131	325	1.01	.12	1.19	4.4	1.30	5.2	.07	.34	60.0	66.1
76	218	325	-.23	.12	1.16	3.2	1.36	4.6	.07	.31	66.5	69.2
78	46	325	2.57	.17	1.09	.9	1.47	2.7	.07	.26	85.5	85.9
59	81	325	1.81	.13	1.14	2.0	1.38	3.7	.07	.31	74.5	76.0
17	294	325	-1.89	.19	1.06	.5	1.19	.9	.09	.19	90.5	90.5
35	196	325	.10	.12	1.14	3.2	1.21	3.6	.14	.33	59.1	66.1
50	102	325	1.45	.13	1.13	2.4	1.18	2.4	.15	.32	68.0	70.8
9	216	325	-.20	.12	1.12	2.5	1.15	2.1	.15	.31	64.6	68.9

Item Statistics

The item statistics tables allow us to check more closely how individual questions have been performing, and how many students have been choosing the correct answers as opposed to the distractors. The tables have two sections that we use. The first, output table 14.1 (Figure 4), shows the number of students who answered the question correctly, the number of students who answered the question, the infit statistics, the outfit statistics, and the point measure correlations.

Figure 4. *BET 2 2019 Output Table 14.1 Item Statistics (cropped for publication)*

ITEM	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%
2	249	325	-.75	.14	.92	-1.2	.84	-1.6	.40	.28	76.6	76.7
3	210	325	-.11	.12	.88	-2.6	.83	-2.7	.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	-.7	.74	-1.6	.36	.22	87.4	87.4
6	65	325	2.12	.15	.96	-.4	.97	-.2	.33	.29	81.8	80.5
7	195	325	.11	.12	1.02	.4	1.02	.4	.30	.33	65.5	66.0
8	268	325	-1.14	.15	.96	-.5	.96	-.3	.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	-2.7	.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	3.5	-.10	.23	89.2	89.5
13	172	325	.44	.12	.90	-2.7	.88	-2.7	.46	.34	70.2	64.5
14	248	325	-.73	.14	.97	-.5	.97	-.3	.32	.28	78.5	76.4
15	211	325	-.12	.12	.88	-2.8	.81	-3.1	.49	.32	71.1	68.1
16	219	325	-.25	.12	1.06	1.2	1.07	1.0	.24	.31	67.4	69.4
17	294	325	-1.89	.19	1.06	.5	1.19	.9	.09	.19	90.5	90.5
18	238	325	-.56	.13	.94	-1.1	.89	-1.2	.38	.29	75.7	73.7

We use output table 14.1 (Figure 4) to review point measure correlations, outfit statistics and infit statistics. For point measure correlations, we want them all to be positive. If they are negative, then the question is not working as expected. Equally, the closer to 0 a score is, the less it is measuring the construct. In Figure 4, we can immediately see that as suggested in item correlation output table 26.1 (Figure 3), questions 12 and 17 are not working as expected. However, if an item is not strongly positive, but all the outfit and infit statistics are within range, it is not necessarily a 'bad' question. Next, we check the outfit statistics, using the MNSQ scores and the Z-scores. The MNSQ scores should be between 0.8 and 1.3. A high score indicates an unpredictable item, which is worse than a predictable low score item. Z-scores should be between −2 and +2. These are the opposite of the MNSQ scores, so here an overly positive score is worse than a negative score. MNSQ takes precedence, so if that score is within range, then Z-scores can be ignored. Z-scores measure the significance of the MNSQ but are more sensitive to sample size. If test takers outside the question's target difficulty are getting this item wrong or right through tiredness or guessing, then these statistics can be thrown off. Questions 12 and 17 appear to be in range but point measure correlation trumps all, and we still also need to look at the infit statistics. Questions 3, 10 and 15 now also need to be looked at. For the infit statistics, we again check to see if the MNSQ and Z-scores are within the same boundaries as the outfit statistics. We are less likely to find problems with infit as it checks that an item is being answered correctly by those test-takers who are predicted to get it right. If an infit statistic is off, the question is measuring something completely different. Clearly question 12 is not working, and questions 3, 10, 15 and 17 at the very least deserve further investigation.

After completing the above checks on all 89 items, we then look at output table 14.3 (Figure 5 and 6) to see how many students are choosing the correct answer versus the distractors. We can also see if the distractors are working in a 'positive' way, i.e. constantly distracting students of lower ability while being ignored by higher ability students. The Data Code refers to the answer choices (A, B or C), and the 1 in the Score Value column indicates the correct answer. The Data Count column refers to the number of students choosing each answer. The average ability indicates the logits of test-takers who chose each option. These should have the lowest at the top and the highest, the correct answer, at the bottom. The PTMEA correlation indicates how well each option is measuring the construct. Only the correct answer should be positive.

Using question 12 as our first example (Figure 5), we can see immediately that only 10% are actually getting it correct, and some of those theoretically should not be. Also, only 2% of students are choosing the other distractor as an option, rendering it useless.

Figure 5. *BET 2 2019 Output Table 14.3 Item Statistics (cropped for publication) and Question 12*

ITEM	DATA CODE	SCORE VALUE	DATA COUNT	%	ABILITY MEAN	P.SD	S.E. MEAN	INFT MNSQ	OUTF MNSQ	PTMA CORR.
12	A	0	7	2	.03	.62	.25	.5	.6	-.11
	B	0	284	87	.61	.76	.05	1.1	1.1	.13
	C	1	34	10	.37*	.66	.12	2.0	1.9	-.10

12 What do you think number 1 is?

A Thank you for telling me.

B I think that's right.

C Let's ask the teacher.

Looking at the question, we can see that distractor A is too obviously wrong, and that distractor B, in certain circumstances could be correct. Option C, the correct answer, while having been taught as part of useful classroom language, has probably not been used by many A2–B1 stream students who would normally just answer the question in class.

Figure 6. *BET 2 2019 Output Table 14.3 Item Statistics for Questions 22, 36 and 31(cropped for publication)*

ITEM	DATA CODE	SCORE VALUE	DATA COUNT	%	ABILITY MEAN	P.SD	S.E. MEAN	INFT MNSQ	OUTF MNSQ	PTMA CORR
22	C	0	20	6	-.24	.41	.09	.7	.6	-.28
	B	0	8	2	-.23	.34	.13	.8	.6	-.17
	A	1	297	91	.65	.73	.04	1.0	1.0	.33
36	A	0	13	4	.37	.46	.13	1.0	.9	-.06
	C	0	139	43	.46	.76	.06	1.2	1.2	-.13
	B	1	173	53	.68	.75	.06	1.1	1.1	.15
31		0	1	0	-.74	.00		.3	.3	-.10
	A	0	51	16	.13	.59	.08	.9	.8	-.26
	B	0	66	20	.42	.61	.08	1.1	1.1	-.11
	C	1	207	64	.74	.77	.05	1.0	1.0	.29

As can be seen above in Figure 6, question 22 is working correctly as a question, but as indicated in output table 1.2 (Figure 2), it is very easy, with 297 students (91%) getting it correct. In question 36 almost as many students choose distractor C as choose the correct answer B. This question is of a type where students read a passage and then decide if the following statements are 'A) Right, B) Wrong, or C) Doesn't Say'. It appears that students can see that the statement is not 'Right' but are confused as to whether it is 'Wrong' or 'Doesn't Say'. Question 31, despite being skipped by one student, is a good example of a reliable item with correctly functioning distractors.

Any items that have been flagged during the Rasch analysis are also double-checked against the results of the Excel Analysis Database outlined below before being re-written.

BET Excel Analysis

A second piece of the GEAC's BET analysis comes from Microsoft Excel. In addition to its role in facilitating students' individual test results and grades, the Excel BET database analyzes the correct answer percentage for each question. These calculations are done individually per class as well as amalgamated by course, where averages for the A1–A2 low stream are separated from the A2–B1 high stream. Figure 7 demonstrates the database's question calculations for a selection of the 2019 April BET 1.

Figure 7. *April 2019 BERT 1-3 Question Totals*

Questions		Totals				A1-A2 Stream Classes							A2-B1 Stream 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-B1	50%	70%	74%	62%	86%	64%	67%	87%	100%	100%	97%	100%	97%
BERT 1	6	36%	23%	46%	A2-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-B1	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-B1	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%

Once these calculations are completed, the correct answer percentages for each question, including overall and course streamed figures, from the cohort's three BETs are put together into a separate Excel database. There, the questions are sorted by difficulty, with result flags given for the ten most difficult questions and the ten easiest questions by test. Figure 8 demonstrates this process.

The correct answer percentage difference between the combined high streams (A2–B1) and the low stream (A1–A2 classes) are also calculated and ranked to determine the top ten of each. The ten questions with the highest gap in answer percentages are labeled as *Ability Determiners*, meaning that these questions are the most responsible for the higher scores of the high stream students. Furthermore, once these questions are examined individually, it is possible they may point to the specific skills and vocabulary recognition that higher stream students tend to possess over their low stream counterparts.

Figure 8. *April 2019 BERT 1-3 Excel Question Analysis*

Test Info		BET 1 (April 2019)							
Section	Question	Overall	Low (A1-A2) Stream %	High (A2-B1) Stream %	Easiest?	Hardest?	Stream % Gap (High - Low)	Outlier? [Low stream higher or close to High stream]	Ability Determiner? [High stream highest over Low stream]
BERT Part 1		**64.2%**	**49.9%**	**75.4%**			**25.4%**		
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		**55.0%**	**39.1%**	**67.5%**			**28.5%**		
BERT Part 2	7	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		**47.3%**	**36.6%**	**55.7%**			**19.1%**		
BERT Part 3	12	11%	10%	11%		1st hardest	1%	T-5th biggest outlier	
BERT Part 3	13	42%	32%	50%			18%		
BERT Part 3	14	36%	26%	45%			19%		
BERT Part 3	15	77%	66%	86%	8th easiest		20%		
BERT Part 3	16	71%	50%	88%			38%		4th biggest AD

Conversely, the ten questions with the lowest gap between streams are labeled as *Outliers*, as they indicate that low stream students are performing at the same level, or in some cases, better than their high stream counterparts. As with the Ability Determiners these questions need individual examining to understand why they are not functioning as intended, as it is possible that ambiguous, nuanced, or exceedingly difficult ideas in the question or testlet are leading higher ability students toward a particular Distractor while conversely rewarding lower-level students for guessing without fully understanding the question. Specifically with BET 1, students take the test before they are identified as low or high stream students and the labels are applied retroactively. This means Outlier questions potentially limited their otherwise superior performance, leading to potential confusion about actual ability levels for course streaming.

2019 BET Excel Microanalysis

This section analyzes individual question examples that exemplify the categories outlined above. The following questions all come from the April 2019 BET 1.

Figure 9. *April 2019 BET 1 Question 26 (cropped for publication)*

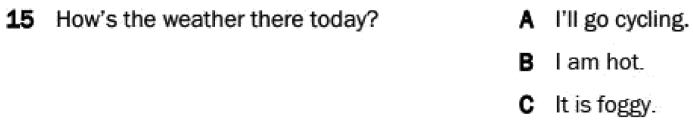
Sayaka met a police officer, and ___25___ police officer asked Sayaka to describe her earrings. She said they were gold and shiny. The officer then asked ___26___ contact information, ___27___ she gave ___28___ her name and her address.

26 **A** for Sayaka to **B** Sayaka for her **C** to Sayaka

Question 26 in Figure 9 above from Reading Part 5 ranked as the 5th hardest among 79 total questions in the 2019 BET 1, with an overall correct answer rate of 22%. However, the answer percentages among the streams split as expected, 16% to 27% for the low and high streams, respectively. While roughly half of both streams incorrectly selected A (48% vs 52% low/high), Distractor C was chosen considerably less by high stream students (35% vs 21% low/high). The

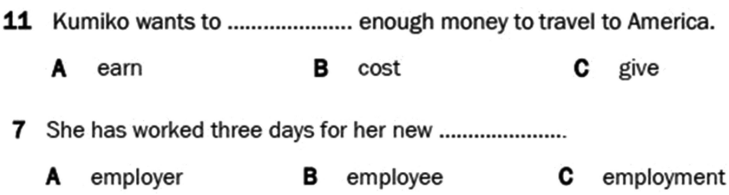
numbers here demonstrate that while most students struggled recognizing *contact* not as a verb but as part of a noun (with *information*) and thus failing to correctly identify the need for an immediate direct object, higher ability students predictably were more able to correctly do so and also had less trouble avoiding the less-grammatically sensical Distractor C. Thus, this question, while difficult, seems to be working as intended to stream students' levels.

Figure 10. *April 2019 BET 1 Question 15*



Similarly, Question 15 (Figure 10) from Reading Part 3 also works solidly to stream students, albeit at the opposite end of the difficulty spectrum. With a correct answer rate of 77%, this question saw a 66% to 86% correct answer split among low to high streams. High stream students were four times as likely (16% vs 4% low/high) to avoid the context-irrelevant Distractor A and nearly twice as likely to avoid Distractor B (18% to 10% low/high), which contained the simpler context-relevant word *hot* but an answer pattern inconsistent with the question. Thus, the data shows that higher ability students could more consistently identify both the correct context-relevant *foggy* and the correct answer pattern.

Figure 11. *April 2019 BET 1 Questions 11 and 7 (cropped for publication)*

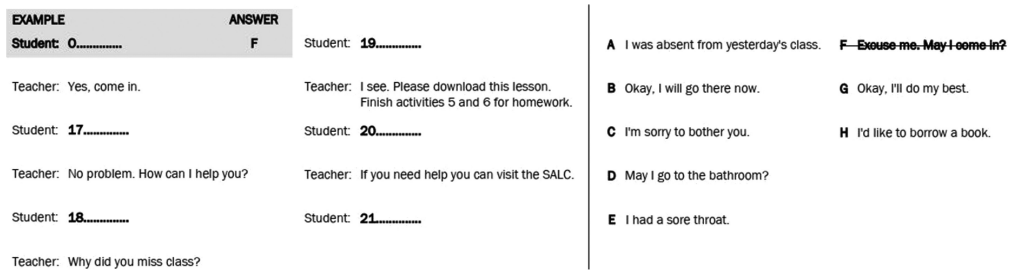


Questions 11 and 7 in Figure 11, despite coming from the same testlet (Reading Part 2), present one of the biggest Ability Determiners and Outliers, respectively, in the 2019 BET 1. Question 11 saw a 53% correct answer percentage gap (18% vs 71%) between low- and high-level streams. On this question, which centers on the vocabulary word *earn* toward being able to afford a trip abroad, low stream students selected Distractors B (*cost*) and C (*give*) at a 45% and 37% rate, respectively. This demonstrates that these students largely did not know the A2 level word *earn* nor associate it with money, instead selecting either the other word with a money context *cost* or assuming that *give enough money* equated to being able to afford something. Conversely, seven out of ten high stream students had a sufficient understanding of *earn* to be able to correctly select it in this context. Thus, the Excel data indicates that this question is highly successful in sorting students' ability levels.

On the other hand, Question 7 shows a reversal of expected outcomes, where low stream students slightly outperformed high stream students (43% vs 41%). Here, the Distractor selection splits were largely identical across both streams, meaning that neither stream had the upper hand at distinguishing between the correct suffix needed among these three B1 level words. Thus, the Excel data shows that this question was not an appropriate indicator of students' actual levels in

this exam and needs to be monitored in the future and or rewritten to allow for a clearer delineation of student ability.

Figure 12. *April 2019 BET 1 Listening Part 4 (cropped for publication)*

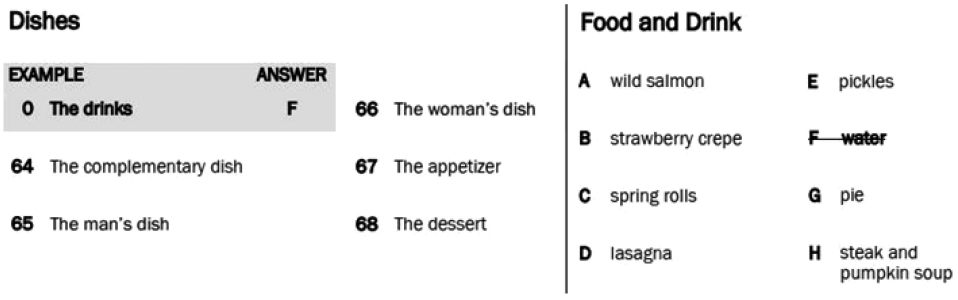


Question 19 (Reading Part 4), as shown in Figure 12, also resulted as one of the largest Ability Determiners in this exam, although the entire testlet needs examining to determine the reason due to its connected conversation chain. Sixty-one percent of high stream students could correctly identify answer E as the answer to Question 19 as opposed to only 27% of low stream students, a gap of 34%. Low stream students instead were drawn heavily (45%) to Distractor A for Question 19, meaning that although they understood the context of a student being absent, they were unable to process the teacher's "Why did you miss class?" was a response to Distractor A (Question 18) rather than a prompt for it. Among students who correctly answered A to Question 18 (43% vs 68% low/high), less than half of the low stream students (47%) were able to follow up the teacher's question with the contextually correct reason *sore throat*, with Distractors C and H accounting for 19% and 20% respectively of the chosen follow-up responses. Conversely, among high stream students who correctly identified answer A for Question 18, 83% of them were able to follow up with the correct response for Question 19.

Within this testlet, Distractor H proved popular among both streams, particularly with the high stream students for Questions 17 and 18, as the context of visiting a teacher's office to borrow a book may have seemed a natural choice despite its lack of a linguistic fit. On the other hand, high stream students largely ignored Distractor D about visiting the bathroom, selecting it only 17 times across the testlet (13 as the answer to Question 17) while low stream students selected it 79 times (54 as the answer to Question 17), meaning that while high stream students were mostly able to eliminate this choice due to its inappropriate context, low-level students were less apt at reading the conversation and context as a whole and thus more likely to choose each response in a vacuum. Additionally, low-level students may have become confused about the setting of the conversation and assumed it took place in a classroom despite the instructions indicating it occurs in a teacher's office.

The final example is Question 66 (Figure 13), coming from Listening Part 2. In this testlet, students must correctly identify the dish that corresponds to each portion of the meal. Only 12% of low-level students and 10% of high-level students correctly identified the lasagna as the woman's dish, making it the third-largest Outlier in the exam. However, the question is written in a way

Figure 13. *April 2019 BET 1 Listening Part 2 (cropped for publication)*



Dishes

EXAMPLE		ANSWER		
0	The drinks	F	66	The woman's dish
64	The complementary dish		67	The appetizer
65	The man's dish		68	The dessert

Food and Drink

A	wild salmon	E	pickles
B	strawberry crepe	F	~~water~~
C	spring rolls	G	pie
D	lasagna	H	steak and pumpkin soup

that on reflection, could be considered to be deliberately deceitful: the woman expresses a desire for salmon (Distractor A), only to be informed that it is sold out, and resignedly selects "Today's Special" as an alternative, which was only identified as lasagna earlier in the listening. Nonetheless, Distractor A was selected 64% to 43% by high vs low-level students as the answer to Question 66, indicating that their increased comprehension level of the text actually led them astray for failing to comprehend the contextual caveat in the sentences that followed. It is possible that this question may be too difficult at the A2 level and demonstrates the capacity to which higher ability students are capable: while they were able to pick up the initial answer, they were largely unable to retroactively apply new information to the answer when the situation changes. Furthermore, the spelling of *Lasagna* is both not how it sounds and quite probably unfamiliar to our students, potentially causing additional difficulties. Thus, this difficulty level ends up rewarding lower ability students for their *inability* to catch the initial food request, and to not be deceived by the information change, necessitating the question be monitored in the future.

BET 2019 Excel Macro Analysis

Having examined examples and possible reasons for questions as Outliers or Ability Determiners, let us now turn to examining the overall data across the three exams.

Figure 14. *2019 BET Overall Section Totals*

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

Figure 14 shows the section difficulty, stream performance gaps, and overall Ability Determiner ranking of the fourteen BET sections across all three 2019 BET exams. Overall answer percentages fluctuated between 65.6% for Reading Part 1 to 43.7% for Reading Part 7. Reading Part 7 also saw the widest gap between courses, with high stream students garnering 27.7% more correct answers than low stream students. Listening Part 6 showed the smallest gap between streams at 11.0%. Overall, reading sections accounted for seven of the top eight largest ability gaps between the streams, with the lowest three ability gaps and five of the lowest six arising from the Listening sections. Building on Figure 14, the top seven testlets with high to low stream ability gaps were all reading testlets, with Reading Part 7 from all three BETs being represented in the top ten (2nd, 3rd, and 10th overall). In contrast, listening testlets make up the lowest six and eight of the ten lowest ability gaps, with Listening Part 6 from the BETs 1 and 3 taking the top two spots (Listening Part 6 from the BET 2 ranked 12th weakest). While the reason for this discrepancy is not immediately clear, it seems that the overall higher difficulty of the listening section resulted in lower scores for both streams and thus a smaller gap between them. The fact that the Listening section follows the 45-minute Reading section and thus is susceptible to student fatigue should also not be discounted.

Figure 15. *2019 BET Categorized High Vs. Low Stream Answer Percentage Gaps*

Testlet	Negative	0% to 5%	5% to 10%	10% to 20%	20% to 30%	30% or more	Average	Rank
Reading 1	0%	7%	7%	40%	40%	7%	**19.7%**	**7**
Reading 2	7%	0%	7%	27%	33%	27%	**22.4%**	**3**
Reading 3	7%	13%	7%	27%	7%	40%	**20.6%**	**5**
Reading 4	7%	0%	0%	13%	40%	40%	**26.0%**	**2**
Reading 5	0%	0%	10%	47%	27%	17%	**21.0%**	**4**
Reading 6	0%	14%	10%	33%	29%	14%	**17.9%**	**8**
Reading 7	0%	0%	13%	13%	20%	53%	**27.7%**	**1**
Reading 8	13%	3%	3%	40%	33%	7%	**16.4%**	**11**
Listening 1*	10%	5%	20%	30%	30%	5%	**15.8%**	**12**
Listening 2	7%	7%	7%	40%	33%	7%	**17.9%**	**9**
Listening 3	0%	0%	13%	33%	27%	27%	**20.5%**	**6**
Listening 4	0%	7%	20%	53%	13%	7%	**14.4%**	**13**
Listening 5	0%	11%	17%	39%	22%	11%	**17.1%**	**10**
Listening 6	6%	11%	28%	44%	11%	0%	**11.0%**	**14**

**BET1 Listening 1 Excluded*

Figure 15 breaks down the fourteen testlets into answer percentage gap categories. This data indicates that the tests are not uniform across each section, with individual questions within each testlet requiring differentiated skill sets despite being designed at the same ability level. Eleven questions resulted in a negative correlation between a correct answer and ability, while fourteen more saw a 5% ability gap or less. Thus, the data demonstrates more attention must be placed on standardizing the difficulty and skills required within each testlet.

With the exception of Reading Part 7, testlets modeled after the PET (Reading parts 7 and 8, Listening Parts 1, 5, and 6) resulted in lower ability gaps. This may indicate that KET style questions, being more straightforward and less reliant on inferences or information across multiple sentences, provide a better indicator as to what separates a higher ability Hiroshima Bunkyo University student from a lower one, while PET questions, being higher in difficulty and thus further beyond the high-to-low spectrum, tend to blur that distinction. Going further, Reading

Part 8 and Listening Part 6, the testlets intended as most difficult per section, resulted in the smallest ability gaps for each respective section across the three exams, with nearly half of Listening Part 6's questions ranking 10% or less. However, these sections are the only testlets with only two answer choices per question, and as Excel cannot determine random guessing from ability, it is possible students' true ability is being obscured. In contrast, Reading Part 7 has four answer choices per question, which significantly reduces the reward for random guessing. As such, over half of this testlet's questions saw an ability gap of 30% or higher. This indicates that difficult testlets with more answer choices may provide more clarity into the B1 spectrum of student ability than testlets with only two answer choices. If implemented, while test scores overall would likely drop, the data portends that such sections may be a better indicator of actual student ability and thus be a better use of difficult test sections.

BET Text Analysis

As we looked at updating and improving BET test items based on Rasch and Excel Database analyses, we realized that while we were writing BET texts based on the CEFR specifications and aims for each test section, the General English curriculum, and using CEFR level-appropriate vocabulary, we did not know exactly how accurately the texts aligned with these criteria. This was an important issue to overcome, since the Rasch analysis outlined at the beginning of this paper only identifies questions that may be too easy or difficult, not why.

We hoped that using Text inspector (TI) software would give us a clearer picture of whether our texts were the appropriate level or not, and if not, what we could do to improve them with the goal of accurately assessing our students. While CEFR analyzes the passage on the individual-word level, Flesch-Kincaid readability test results give us an approximate overall reading difficulty score. Therefore, in the spring semester of 2019, we carried out a TI analysis of all reading texts and listening scripts for all three BETs. The results included a CEFR score for all of the vocabulary used in each text and script, as well as Flesch-Kincaid readability test scores for each part.

Flesch-Kincaid readability tests are popular methods used to assess readability of texts. They are used in a wide variety of disciplines and are designed to indicate how difficult a written passage is to understand. We used both tests in our BET analysis: the Flesch Reading Ease (FRE), and the Flesch-Kincaid Grade Level (FKL). They both use the same core measures, but they correlate inversely: A higher score on the FRE should have a lower score on the FKL. With ASL equaling average sentence length and ASW equaling average number of syllables per word, the FRE formula is as follows:

$$FRE = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

In the FRE, higher scores indicate that the text is easier to read; conversely, lower numbers indicate that the text is more difficult to read. While there are technically higher and lower scores

possible, most scores fall in the range of 0-100. 100 is considered to be a 5th-grade reading level in the U.S., while 0 would be a professional or academic level. (Linney, 2020)

The FKL formula is the following:

$$FKL = (0.39 \times ASL) + (11.8 \times ASW) - 15.59$$

Therefore, a lower FKL indicates a passage is easier to read. In practical terms most scores are between -1 and 18, where -1 would indicate sentences mostly made of one-syllable words. Eighteen would indicate an academic-level text. (Linney, 2020)

What is Text Inspector?

Text Inspector (TI) is an online tool for analyzing and measuring the vocabulary and discourse difficulty level of English texts. We used the TI on www.textinspector.com. On the subscription-based TI website, you paste in a selection you would like to have analyzed, and a number of metrics are used to evaluate your text. In the fall semester of 2019, we analyzed all three of our BETs, which included all eight reading parts (BERTs) and transcripts of all six listening parts (BELTs) for each BET using the three metrics mentioned above: CEFR score, FRE, and FKL. Figure 16 shows an example output.

Figure 16. *Screenshots of the Text Inspector analysis of BELT 3 part 6*

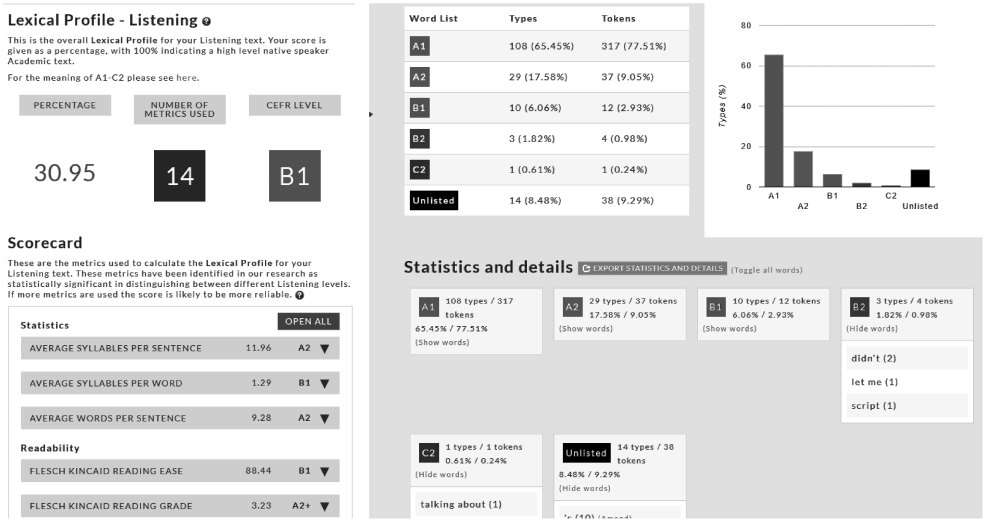


Table 1 shows the results of the TI analysis for the reading section of BET 1. The Text Inspector software accuracy varied based on the length of texts: longer texts provided more accurate results. Therefore, results were not provided for some of the shorter reading sections since the software could not guarantee accuracy.

Table 1. *BET 1 Reading Section (BERT 1) Text Inspector Analysis Terminology and Results*

Term	Meaning	Term	Meaning
CEFR Aim	The CEFR level goal as outlined in the BET specifications for each section of the test.	FRE/ FKL	The Flesch Reading Ease score/Flesch-Kinkaid Grade Level given by the Text Inspector
Vocab CEFR	The CEFR level of each word in the text.	NA	Designation for data that was not available.
Overall CEFR	The overall score given by the Text Inspector		

		Vocabulary CEFR					
Part	CEFR Aim	A1	A2	B1	B2	Overall CEFR	FER/FKL
1	A2	73%	27%	–	–	NA	NA
2	A2/ A2+	61%	18%	15%	–	NA	NA
3	A2/ A2+	69%	7%	5%	–	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

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+. Vocabulary CEFR percentages will not equal exactly 100 because there is some unlisted vocabulary (for example, specific vocabulary related to Bunkyo University and the BECC). Also, there were a few words that were listed by the TI that were higher than B1+ but after discussion, were not included because either they were taught during the semester and therefore would be familiar to students (for example, *campus*), or they were considered to be easier than the TI specified (for example, *talked about*).

The TI analysis found that overall, the GEAC has been successful in creating reading passages and listening scripts that fit the overall CEFR Aims of each section of the BET test. However, some BERT parts skewed slightly easier than their specified Aims. For example, Overall CEFR for BERT 1 part 4 was rated as A1+ but the CEFR Aims suggest A2/A2+.

Also shown in Table 1, FRE/FKL scores show the BETs as gradually becoming more challenging in later parts. For example, BERT 1 part 3 had an FRE score of 90 and an FKL score of 1.8, which equate to a 5th grade reading level in the U.S. The most difficult part, BERT 1 part 8, had an FRE score of 65 and an FKL score of 8.5, which equate to an 8th or 9th grade reading level in the U.S. As we described earlier, there should be an inverse relationship between FRE and FKL, and indeed we observed that relationship throughout all the BETs results. As with the Overall CEFR, FRE/KFL scores were not available for some parts of the BET because the text or scripts were

too short, and therefore the scores were deemed unreliable by the TI.

Though the Overall CEFR level was not available for BERT part 2 and 3, the Vocab CEFR analysis showed that the percentage of A2 and B1 vocabulary was higher in BERT part 2 than BERT 3. This suggests that BERT part 2 was actually more difficult than BERT part 3, despite the correct answer percentage found in the Excel analysis. Also, the TI analysis for BERT part 6 showed that it was more difficult than BERT part 7.

Table 2. *BET 2 Listening Section (BELT 2) Text Inspector Analysis Results*

		Vocabulary CEFR					
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
5	B1/B1+	57%	21%	12%	2%	B2	83/ 3.85
6	B1/B1+	70%	15%	7%	1%	B1	83/ 4.26

The results of the TI of the BET listening section as seen in Table 2 above shows the CEFR Aims to be similar to those in the reading section. Also, the GEAC committee designed each part to generally increase in difficulty. The same caveat mentioned earlier also applies to the BELT Vocab CEFR results: percentages will not equal 100, because there are always some unlisted words that appear in our vocabulary list related to the GE curriculum and a few special cases of words listed as higher than B2 by the TI (for example, words related to school life).

However, in the BELTs, and adding the findings produced by the earlier mentioned Excel analysis, overall CEFR showed that this section was generally more difficult than the CEFR Aims outlined, which was not the case in the BERTs. This probably has to do with the fact that we were analyzing listening scripts, rather than reading texts. Since listening tasks are generally considered easier to understand and answer than reading tasks, it would be natural that the CEFR levels were more advanced. We found similar results for the FRE/FKL scores, in which the overall reading ease ranged between 86/3.0 points and 83/4.26 points (all considered within US Grade 6 reading level). While the lack of progression in difficulty in FRE/FKL scores could be a concern, it is important to remember that these are scripts, and therefore do not take into consideration the speed nor accent in which these scripts are read. We deliberately use a mix of American, British, New Zealand, Philippine and Japanese voices in our class materials and in the BETs. Unfortunately, TI does not have a way to rate these factors.

Based on the results of the TI analysis, and in conjunction with the Rasch and Excel analyses, we decided to make two changes to the BET. First, we switched the order of BERT part 2 and 3, based on the TI findings discussed earlier in which BET part 2 was found to be more difficult than BET part 3. Next, we removed BERT part 6 since TI and Rasch analyses showed that it was a

redundant section and not testing unique language skills.

Overall, the Text Analyses of the Bunkyo English Tests showed us that the revisions undertaken every year over the past six years have resulted in assessments that meet the CEFR Aims of the assessments. The CEFR Vocab, Overall CEFR, and the FRE/FKL scores were in general agreement. Revisions and updates will continue, and the details of revisions undertaken after the TI analyses as well as the RASCH analyses will be discussed in the next section. In the future, any new reading or listening parts that are added to the BETs will be analyzed using TI and modified if they deviate significantly from the CEFR Aims.

Concluding Comments

After performing the analyses listed above, the question still remains: Are our BETs valid? Shaw, S (2020) quotes Standards (APA, AERA, NCME, 1966) as saying that "it is incorrect to use the unqualified phrase 'the validity of the test.' No test is valid for all purposes or in all situations or for all groups of individuals". All we can do is to try to provide evidence of having performed due process to satisfy the five earlier mentioned stages of construct validity:

- Does the content of the test match the content of the curriculum?

While the lesson handouts have gone away in favour of iPad-based lessons, we still base our BET questions on, and regularly check them against the curriculum lesson contents and our vocabulary list. At the same time, in order to answer Shaw's first question regarding content validity, we take several steps. To avoid the issue of students answering incorrectly because they did not understand what to do, rather than not having the English ability, all instructions in the BETs are in Japanese. Also, we make sure the items cannot be answered through rote memory. Pitts and Naumenko (2016) state that:

"It is thus left to a teacher, as a professional within his/her content area and grade level, to determine what constitutes "opportunity to learn" without artificially inflating student scores through test preparation activities like "drill and kill" item exercises. Specific guidelines are not available to direct decisions about instructional practices so as to ensure that students are aware of the test domain, nature of items, mastery criteria and modes of test administration while avoiding the artificial inflation of test scores through inappropriate test preparation activities."

In many cases, this lack of guidelines and onus on the teachers when dealing with tests based on content validity causes teachers to naturally lean towards 'teaching for the test', rather than letting the assessment measure what the students have been studying. In our case, to combat this, no questions in the BETs are exact copies of anything in lessons. We write question items based on the topics, notions, functions, situations, and vocabulary used throughout the curriculum, but students and teachers never see any of the questions. While ensuring a fair playing field for all students, this is still not without issues. An example is the previously mentioned 2019 BET 2

Question 12. While taken directly from materials used in class, it is failing to function as a useful question. As the curriculum and its vocabulary change, we will continue to try to match the content, but in a way that reflects its most widespread usage.

- Do students answer the questions in the manner intended, and do the students' answers to different questions relate in a way we would expect them to?

As previously mentioned in the BET Rasch and BET Excel Analysis sections, we can identify that many of our questions are being answered as intended and that students answer consistently. By using these two methods together however, we can also more readily pick out questions or distractors that are not being answered as we would expect. While not emphatically telling us 'why' a student or set of students get certain questions unexpectedly wrong, the data can certainly lead us toward more focused discussions. These can be around whether the question should be changed in terms of vocabulary used or overall complexity, or if the lesson materials and teaching emphasis should be looked at.

- Are the tests marked in the way they are intended to be?

In our BETs, all items carry the same weighting within the test, which is then graded by machine. In this way, we can remove the human error element of grading and ensure fair scores are given to each student.

- Can users of the results interpret them in the manner intended?

The initial aims of our BETs were, and remain, threefold: To stream, to monitor progress, and to give informative feedback to students in terms of a widely recognized scale of English language ability. From a 'teachers as users' perspective, while "Single test score-based decisions are inherently inappropriate as they are based on an insufficient summary of test taker achievement" (Pitts and Naumenko, 2016), like many institutions, we are given no other evidence with which to initially stream our GE students. For us, "The overarching goal of streaming is to place students as fairly and evenly as possible into the class sections within the three different sub-streams while meeting all of the BECC's ideal streaming criteria" (Svien, 2019). Through the use of Rasch fit statistics and Excel ability determiner results, we hope to be as accurate as possible so that students find themselves working with the correct levels of materials. Coupled with the current end of semester speaking assessments, and soon to be unit based speaking assessments, we hope to further improve the accuracy of our streaming for second-year students.

At the same time, improved reliability in results will mean that we can provide more accurate feedback to students and their teachers as to overall (or lack of) student progress. From now on a major part of this will be looking at the overall CEFR level of individual parts of the test, in conjunction with the results produced by the Excel analysis. How each part is written and performs washes back strongly into the curriculum as the majority of classroom reading and

listening activities are based on the specifications for writing the BETs. As mentioned earlier, switching the order of BERT part 2 and 3, and removing BERT part 6 (the Right, Wrong, Doesn't Say example mentioned in Figure 6.) will translate to switches in the order of, and removal of similar activities in lessons. This should in turn lead to more level appropriate and level determining materials appearing in our lessons.

If we can improve the CEFR accuracy of individual parts of the BETs, we will be able to give students a better idea of their CEFR levels, as we currently do with in-class assessments. By informing students of their CEFR level performance in reading and listening activities, we can more accurately inform them as to the level of activities and extra work they can be choosing to do in our Self-Access Learning Center, and as to what level and type of external English exams (EIKEN, TOEIC, TOEFL, IELTS etc.) they can choose as most beneficial for their future employment needs.

Sadly, with most of our students living in rural Japan, and having limited access to study abroad programs or international exchange activities, it is impossible to know if they can then actually use the English they have 'passed' in an English test. Only by observing student performance within their classes, and by adding regular in-class practical assessments can we have any real idea of whether or not students can perform at a level that their tests say they can. For these reasons, along with the necessarily narrow nature of our curriculum and the fact that students no longer have to do a compulsory two years of study, the original idea that "students will also receive a CEFR certificate at the end of their study" (Bower, et al, 2014) has had to be reconsidered. However, as we look to encourage more Education and Global Communication students to take extra reading, writing, third, and fourth-year courses, more accurately CEFR levelled reading and listening texts that are proven to be ability determining may yet help this become a reality.

This report aims to detail the processes the GEAC went through when analyzing the 2019 BETs, before rewriting items for the 2020 – 2021 calendar year. Until we have completed all BETs for the year it is difficult to determine if we have made any great improvements in test reliability, regardless of the attempts we have made to try to prove our construct validity. At the time of writing, due to COVID-19 issues, all assessment for the year is under review. Whatever form the January and possibly April 2021 BETs take, we shall continue to use our three-stage Rasch, Excel and Text Inspector analysis process to provide the most relevant and practical form of assessment for our students.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, (AERA/APA/NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Bond, T G., Fox, C M. (2001) *Applying the Rasch Model. Fundamental Measurement in the Human Sciences.* Mahwah, New Jersey, Lawrence Erlbaum Associates, Inc.

Bower, J., Rutson-Griffiths, A., Sugg, R. (2014). *Setting and Raising Standards - the Rationale for, and the*

Structure of the Bunkyo English Tests. Bulletin of Hiroshima Bunkyo Women's University, Volume 49.
Linacre, J. M. (2008). Winsteps Multiple-Choice, Rating Scale and Partial Credit Rasch Analysis is available at: https://www.winsteps.com/winsteps.htm
Linney, S. (2017, 2020). *The Flesch Reading Ease and Flesch-Kincaid Grade Level.* The Readable Blog. Available at https://readable.com/blog/the-flesch-reading-ease-and-flesch-kincaid-grade-level/
Pitts, R T., Naumenko, O. (2016) *The 2014 Standards for Educational and Psychological Testing: What teachers Initially Need to Know.* Working Papers in Education, Vol.2, No.1. University of North Carolina, Greensboro School of Education. Retrieved from http://libjournal.uncg.edu/wpe/article/view/1316
Shaw, S. (2020) *Unpacking Validity.* Presented at Cambridge Assessment, Triangle Building, Cambridge, UK
Svien, J. (2019). *Streaming University English Courses: Best Practices.* Bulletin of Hiroshima Bunkyo University, Volume 54.
Text Inspector is available at: https://textinspector.com/

—2020年9月24日　受理—