Lessons learned from five years of speaking exam administration

Jordan Svien
jsvien.becc@gmail.com
Hiroshima Bunkyo University

https://doi.org/10.37546/JALTSIG.TEVAL26.1-2

Abstract

From 2015 to 2019, the Bunkyo English Communication Center at Hiroshima Bunkyo University conducted end-of-semester speaking exams called *Bunkyo English Speaking Tests* (BESTs) for all English Communication freshman and sophomore students. During these five years, the Bunkyo English Communication Center learned several test administration best practices. First, in a desire to apply a many-facet Rasch model using the Facets software package (Linacre, 2022a) to provide student fair scores that account for rater leniency and severity, a preventative flaw in the rater schedule was discovered and corrected. Second, the increased complexity of the rater schedule plus a desire to streamline the exam processes necessitated the building of a comprehensive scheduling and testing system in Excel. Finally, the calculation method initially used for converting Rasch measures into student fair scores was based on a faulty assumption and suffered from ambiguity and subjectivity, and a fairer workaround system was discovered and implemented. This paper documents the discovery of these problems and the process of developing and implementing their solutions.

Keywords: examination, assessment, MFRM, Facets

Introduction to the BEST

The Bunkyo English Speaking Tests (BESTs) are CEFR-aligned examinations that comprise the final spoken course grade of the Bunkyo English Communication Center (BECC) at Hiroshima Bunkyo University's English Communication courses, a mandatory course entitled Freshman English (FE) for all first-year students and an optional course entitled Sophomore English (SE) for second-year students. Both FE and SE courses are streamed, with low-level and high-level classes respectively aiming to advance students from the A1 to the A2 CEFR band and the A2 to the B1 CEFR band (COE, 2001, updated 2018). The BESTs are held at the end of each semester, entitled BEST 1 and 2 for FE terms 1 and 2 and BEST 3 and 4 for SE terms 1 and 2. First implemented in 2015 and designed by the BECC's General English Assessment Committee (GEAC), they seek to consistently track and evaluate student speaking performance based on the BECC's in-house English Communication course content (Sugg and Svien, 2018). The exam format is based on the Cambridge KET and PET speaking tests (2016), adhering to a dual-rater system. An interlocutor facilitates the exam and scores students via a holistic rubric, while a non-participatory rater provides scores for the analytic rubric, consisting of scores for grammar and vocabulary (combined), pronunciation, and interactive communication. Like the KET and PET, the exams are conducted in pairs, with students communicating both with the interlocutor and each other across three separate tasks. Students are assigned a score for each category from 1 to 5 (with half points allowed for 3 and 4), each corresponding to a CEFR ability band. Table 1 provides a summary and the Appendix provides the full rubric for each category.

Table 1

BEST scoring overview

CEFR Level	BEST Score	Rubric (+Category)		Judge	Weight
B1 or above	5	Holistic		Interlocutor	40%
A2+	4.5				
A2	4	Analytic	Grammar and Vocabulary	Rater	20%
A1+	3.5				
A1	3		Pronunciation	Rater	20%
Pre-A1	2				
Pre-A1	1		Interactive Communication	Rater	20%

The rater's three analytic scores comprise 60% of the total grade, and the interlocutor's holistic score is doubled to form the final 40%. This 25-point raw score is multiplied by 0.6 to form a final grade out of 15, which comprises 15% of the English Communication course term grade. Prior to each BEST, a mandatory standardization session for all judges is conducted consisting of test rubrics, procedures, and practice scoring videos and discussions. For a full overview of the BEST teacher standardization process as well as the development of the BEST rating scale and the specific tasks conducted and assessed, see Sugg and Svien (2018).

Through the summer of 2016, the 15-point converted score was utilized as the students' exam grade. However, beginning in semester 2 of 2016, the "final" piece of this grading process began to be explored: many-facet Rasch measurement (MFRM) conducted via Facets (Linacre, 2022a), a software program for many-facet Rasch measurement. If possible, Facets would correct teacher leniency and strictness that had yet to be ironed out after the standardization sessions. However, it was ultimately several years before this system moved into its final iteration. Over these years, the BECC learned three important lessons: how to successfully build a Facets compatible rater schedule, how to best facilitate the scheduling and roster input process, and how to best process the Rasch analysis results.

Lesson 1: Developing a Rasch-Facets compatible rater system

The BESTs are scheduled across four days of the end-of-semester exam week, with FE and SE courses both holding two days of exams. Students are assigned to one of the two exam days. Facets requires the judging plan to contain sufficient linkage between the elements of all the facets, where "every element can be compared directly and unambiguously with every other element" (Linacre, 1997). With each judge assigning only one or three non-overlapping scores to each student, two questions remained for the GEAC: was there a judging setup which provided enough inter-facet linkage to provide a cohesive frame of reference, and would the amount of data that needed to be declared as "missing" (due to those scores not being assigned by judges of the opposite role) cause Facets to be unable to process the results?

To tackle the first question, each BECC teacher was assigned as either a rater or interlocutor for Day 1 of each test (FE and SE), then given the opposite role for Day 2. This was designed to spread interlocutor and rater coverage as well as possible for Facets in addition to the professional development benefit giving all teachers experience in both judging capacities. While this occasionally entailed the same two teachers who previously judged a class together simply reversing roles, judges were predominantly mixed up so that few teachers saw the same "partner" across the same course. Teachers were eligible to repeat a class but with a different role, resulting in a judging plan as shown in Table 2.

Table 2

BEST rater and interlocutor scheduling system, 2015-2016

Day	Period	Class	Interlocutor	Rater	Day	Period	Class	Interlocutor	Rater
1	1	FE1	1	6	2	1	FE1	8	1
		FE2	2	7			FE2	10	5
		FE3	3	8			FE3	9	3
		FE4	4	9			FE4	6	2
		FE5	5	10			FE5	7	4
	2	FE6	1	7		2	FE6	8	4
		FE7	2	10			FE7	10	3
		FE8	3	8			FE8	9	2
		FE9	4	6			FE9	6	1
		FE10	5	9			FE10	7	5

**FE = Freshman English*

As shown, teachers were assigned numbers to track their positioning across the exam week. Teachers 1, 2, 3, 4, and 5 were assigned to group 1 (Day 1 interlocutors / Day 2 raters), while teachers 6, 7, 8, 9 and 10 were group 2 (Day 1 raters / Day 2 interlocutors).

The MFRM model was configured to estimate three facets: Students, Judges, and Items, with Item 1 (the interlocutor score) given double weight. The facet model statements were entered in the Facets specifications as:

Model = ?,?,1,Ratings,2

Model = ?,?,?,Ratings,1

Although the rater and interlocutor are responsible for different scores (see Table 1), MFRM can accommodate missing data, which is represented by # in Table 3.

Table 3

BEST Facets input file scores example (scores fabricated)

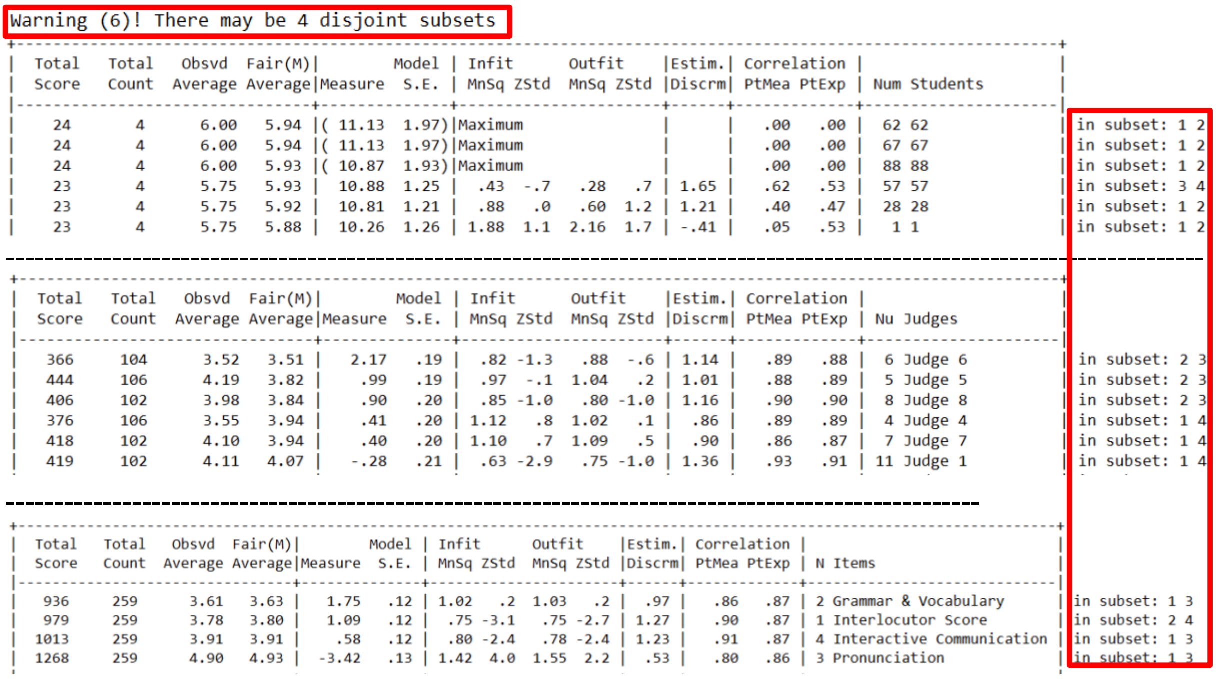
				Interlocutor	Rater		
Day	Student	Teacher	Categories	Holistic Score	Grammar + Vocabulary Score	Pronunciation Score	Interactive Communication Score
1	1	1	1-4a	5	#	#	#
	1	6	1-4a	#	4.5	4.5	5
2	15	8	1-4a	4.5	#	#	#
	15	1	1-4a	#	5	4.5	4.5

Table 3 shows the first and fifteenth students of an example FE1 class based on a model schedule fitting the Table 2 parameters. Student 1 saw Teacher 1 as the interlocutor and Teacher 6 as the rater; Teacher 1 awarded a 5 for the holistic score and Teacher 6 a 4.5, 4.5, and 5 respectively for the rater scores. The scores not assigned by the respective teachers are considered “missing” (# marks). On the second day, student 15 was awarded a 4.5 by Teacher 8 (interlocutor) and a 5, 4.5, and 5 by Teacher 1 (rater).

In this setup, all teachers participate in both judging roles for each course, and teachers rotate through several judging “partners” who provide the opposite role’s score(s) for each student. To begin applying MFRM from the 2016 BEST 2 and 4, the data from the 2015 BEST 2 and 4 and the 2016 BEST 1 and 3 were retroactively modeled using Facets. However, the analysis revealed a flaw in the system that needed identifying and rectifying before MFRM could begin.

Figure 1

2015 BEST 2 Rasch output file subsets



```
Warning (6)! There may be 4 disjoint subsets
+-----------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit     |Estim.| Correlation |                        |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Num Students           |
|--------------------------------+--------------+----------------------+------+-------------+------------------------|
|    24       4      6.00    5.94 |( 11.13  1.97)|Maximum              |      |   .00   .00 |  62 62                 | in subset: 1 2
|    24       4      6.00    5.94 |( 11.13  1.97)|Maximum              |      |   .00   .00 |  67 67                 | in subset: 1 2
|    24       4      6.00    5.93 |( 10.87  1.93)|Maximum              |      |   .00   .00 |  88 88                 | in subset: 1 2
|    23       4      5.75    5.93 |  10.88  1.25 |  .43  -.7   .28   .7 | 1.65 |   .62   .53 |  57 57                 | in subset: 3 4
|    23       4      5.75    5.92 |  10.81  1.21 |  .88   .0   .60  1.2 | 1.21 |   .40   .47 |  28 28                 | in subset: 1 2
|    23       4      5.75    5.88 |  10.26  1.26 | 1.88  1.1  2.16  1.7 | -.41 |   .05   .53 |   1 1                  | in subset: 1 2
---------------------------------------------------------------------------------------------------------------------
+-----------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit     |Estim.| Correlation |                        |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Nu Judges              |
|--------------------------------+--------------+----------------------+------+-------------+------------------------|
|   366     104      3.52    3.51 |   2.17   .19 |  .82 -1.3   .88  -.6 | 1.14 |   .89   .88 |  6 Judge 6             | in subset: 2 3
|   444     106      4.19    3.82 |    .99   .19 |  .97  -.1  1.04   .2 | 1.01 |   .88   .89 |  5 Judge 5             | in subset: 2 3
|   406     102      3.98    3.84 |    .90   .20 |  .85 -1.0   .80 -1.0 | 1.16 |   .90   .90 |  8 Judge 8             | in subset: 2 3
|   376     106      3.55    3.94 |    .41   .20 | 1.12   .8  1.02   .1 |  .86 |   .89   .89 |  4 Judge 4             | in subset: 1 4
|   418     102      4.10    3.94 |    .40   .20 | 1.10   .7  1.09   .5 |  .90 |   .86   .87 |  7 Judge 7             | in subset: 1 4
|   419     102      4.11    4.07 |   -.28   .21 |  .63 -2.9   .75 -1.0 | 1.36 |   .93   .91 | 11 Judge 1             | in subset: 1 4
---------------------------------------------------------------------------------------------------
+-----------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit      Outfit     |Estim.| Correlation |                          |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | N Items                  |
|--------------------------------+--------------+----------------------+------+-------------+--------------------------|
|   936     259      3.61    3.63 |   1.75   .12 | 1.02   .2  1.03   .2 |  .97 |   .86   .87 | 2 Grammar & Vocabulary   |in subset: 1 3
|   979     259      3.78    3.80 |   1.09   .12 |  .75 -3.1   .75 -2.7 | 1.27 |   .90   .87 | 1 Interlocutor Score     |in subset: 2 4
|  1013     259      3.91    3.91 |    .58   .12 |  .80 -2.4   .78 -2.4 | 1.23 |   .91   .87 | 4 Interactive Communication |in subset: 1 3
|  1268     259      4.90    4.93 |  -3.42   .13 | 1.42  4.0  1.55  2.2 |  .53 |   .80   .86 | 3 Pronunciation          |in subset: 1 3
```

As shown in Figure 1, the MFRM data connectivity test failed for both the 2015 BEST 2 and 4, indicating that the three facets had been split into four disjoint subsets, with each item in two of them. Students were placed into subsets 1 and 2 or subsets 3 and 4 depending on the day they took their exam. Conversely, teachers who began as Day 1 raters and switched to interlocutors on Day 2 were placed in subsets 1 and 4, while those with the opposite schedule were put in subsets 2 and 3. Finally, all holistic scores awarded by the interlocutor were placed in subsets 2 and 4 and all rater scores into subsets 1 and 3 (Table 4).

Table 4

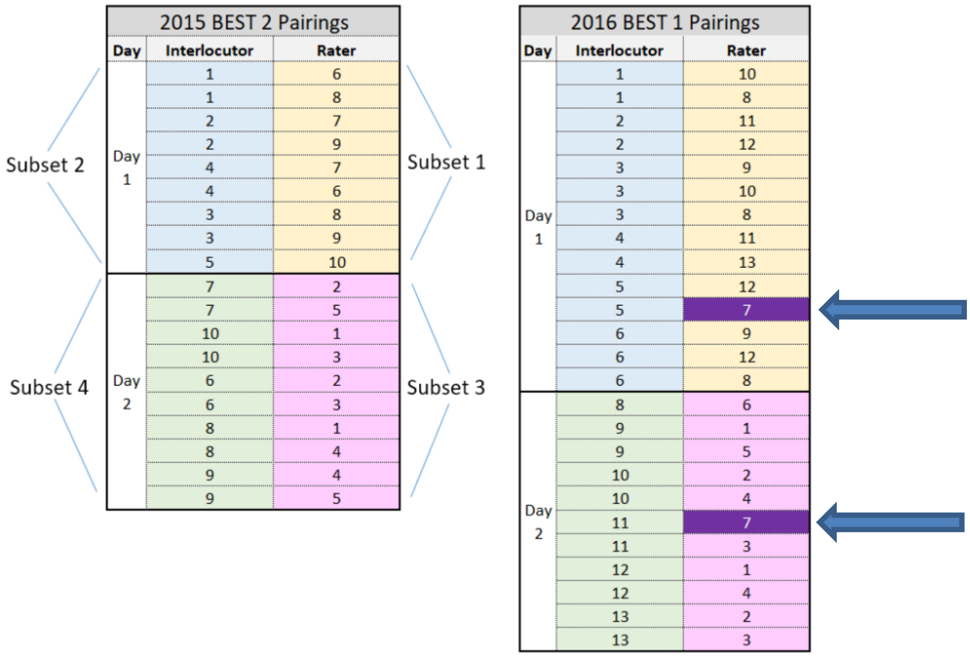
2015 BEST 2 and 4 subset summary

Subset	Students	Teachers	Scores
1	Day 1	Group 1 (Raters)	Rater Scores
2	Day 1	Group 2 (Interlocutors)	Interlocutor Score
3	Day 2	Group 2 (Raters)	Rater Scores
4	Day 2	Group 1 (Interlocutors)	Interlocutor Score

Thus, despite the GEAC's efforts to spread test coverage, linkage between all facets was not achieved. One explanation offered at the time was that Facets was unable to reconcile the data set properly due to the "missing" data on each student score line, and thus it seemed Rasch analysis would not be an option for producing fair scores going forward. Unexpectedly, however, the 2016 BEST 1 and 3 data was *not* divided into subsets, despite being designed with the same judging system as in 2015, indicating that the "missing" data was not the cause of the problem. Rather, the judging system itself seemed to be flawed. Thus, a deeper comparison of what succeeded in the 2016 BEST 1 and 3 but failed in the 2015 BEST 2 and 4 was warranted. Figure 2 below shows a comparison of the 2015 BEST 2 and 2016 BEST 1 from a judging standpoint.

Figure 2

2015 BEST 2 (failure) vs 2016 BEST 1 (success) teacher pairings


Subset 2
Subset 1
Subset 4
Subset 3

2015 BEST 2 Pairings

Day	Interlocutor	Rater
Day 1	1	6
	1	8
	2	7
	2	9
	4	7
	4	6
	3	8
	3	9
	5	10
Day 2	7	2
	7	5
	10	1
	10	3
	6	2
	6	3
	8	1
	8	4
	9	4
	9	5

2016 BEST 1 Pairings

Day	Interlocutor	Rater
Day 1	1	10
	1	8
	2	11
	2	12
	3	9
	3	10
	3	8
	4	11
	4	13
	5	12
	5	7
	6	9
	6	12
	6	8
Day 2	8	6
	9	1
	9	5
	10	2
	10	4
	11	7
	11	3
	12	1
	12	4
	13	2
	13	3

For the 2016 BEST 1, Teachers 1, 2, 3, 4, 5 and 6 comprised group 1 and Teachers 8, 9, 10, 11, 12, and 13 group 2. Teacher number 7, a non-regular testing member who volunteered to “fill-in” the schedule where needed, joined for two total sessions, one per day, in a rater capacity. However, it was discovered that this single discrepancy was responsible for the (tenuous) unification of the data set.

The 4 distinct subsets created in 2015 can be seen in the color coding used in Figure 2. In 2015, the relative severity of each rater can be compared, but only *within* each of the 4 subsets. For example, the average ratings awarded by each rater on Day 1 (Subset 2) and can be ranked from most to least severe by their average ratings. However, they cannot be compared with the average ratings awarded on Day 2 (Subset 4), because both the students who participated and teachers who *rated* on Day 2 were different. The same can also be said about the interlocutor scores on Day 1 and Day 2 (Subsets 1 and 3). In other words, Facets cannot determine whether the students on Days 1 and 2 differed slightly in their ability, whether the teachers who gave ratings or holistic scores on Day 1 and 2 differed in their severity, or whether the items—analytic versus holistic scores—differed in their difficulty.

Rater 7 in the 2016 BEST 1, however, inadvertently provided a means to make those comparisons. Rater 7 was unique in awarding analytic ratings (as opposed to the holistic score) to students on both days. It is a very tenuous connection, but Rater 7’s average ratings can now be used to infer whether the students on Day 1 and Day 2 differed slightly in their ability. More importantly, the average ratings of all teachers can now be ranked from most to least severe by comparing their average ratings to Rater 7. Although Rater 7 never participated as an interlocutor, Facets can use indirect comparisons to rank the interlocutors as well. For example, once it is determined from Rater 7’s analytic ratings whether the students on Day 1 differ from Day 2, it can also be inferred whether getting a high score on the analytic ratings is more difficult than getting a high score from the interlocutor. From there, teacher severity when functioning as an interlocutor can be determined and ranked. In other words, the presence of Rater 7 made it possible for the Facets software to compare and rank the elements of all three facets—participants, raters, and items—and place them on a single logit scale.

Thus, to rectify this error, a new set of criteria was implemented from the 2016 BEST 2 and 4. As the facet linkage in the 2016 BEST 1 was extremely tenuous, a new system to make data linkages an integral component was designed. Rather than two non-overlapping judging groups, teachers are assigned to one of four judging groups for each exam:

As shown in Table 5, while some interlocutor and raters swap roles after each test day, others remain in their roles throughout the exam, guaranteeing internal data connections among these four groups. Furthermore, even if a teacher is absent and a replacement needs to be found, there is no concern over data connection lapses.

Table 5

Role groups for 2016 BEST 2 and 4 onward

Role	FE BEST 1/3 Role		SE BEST 2/4 Role	
Group	Day 1	Day 2	Day 1	Day 2
1	Interlocutor	Interlocutor	Rater	Rater
2	Interlocutor	Rater	Rater	Interlocutor
3	Rater	Interlocutor	Interlocutor	Rater
4	Rater	Rater	Interlocutor	Interlocutor

To further promote data connectivity and exam integrity, the following scheduling rules were added:

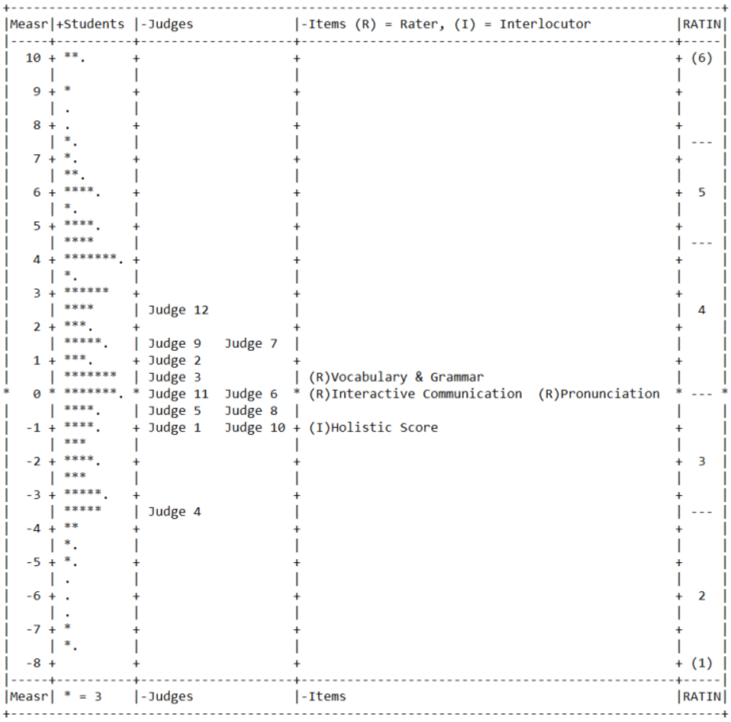
- Raters and interlocutors are not paired together more than once per test.
- To strengthen the integrity of the MFRM and provide as much data on teacher leniency and strictness as possible, teachers are separated after one test together (even if judging roles were to be reversed).
- Teachers do not judge the same class both days.
- This was implemented toward promoting fairness in case of lenient or strict scorers, so that an entire class is not judged by the same teacher. Although Rasch fair scores are used to even out these discrepancies, care is taken to minimize them on the front end.
- Teachers have an even distribution of low-level and high-level classes.

These parameters allow teachers to see a range of student abilities across their testing sessions to better understand the scoring levels. Class levels are not outwardly shared with teachers so that they remain unbiased during the session, but by providing a varying set of levels each teacher's leniency or severity can be more transparent. Teachers whenever possible are not assigned to classes of Global Communication Department (GCD) students they teach in other subjects because students in this department take several other BECC courses. Although teachers complete the standardization session and are required to remain impartial, it is impossible to fully discard any preconceptions of student ability based on their performance in other classes. Furthermore, these students may have an advantage or disadvantage compared to their peers. Some students may be relaxed by the added level of familiarity with the teacher, while others may become more anxious.

Instituting the above procedures eliminated the disjoint subsets, making it possible to compare students, judges, and items on a common scale. The results can be seen most clearly in the Facets Ruler, a visual tool created by Facets that illustrates the relationships between all elements specified in the MFRM analysis (Figure 3).

Figure 3

2018 BEST 1 ruler



```
+------------------------------------------------------------------------------------------+
|Measr|+Students |-Judges            |-Items (R) = Rater, (I) = Interlocutor         |RATIN|
|-----+----------+-------------------+-----------------------------------------------+-----|
|  10 + **.      +                   +                                               + (6) |
|     |          |                   |                                               |     |
|   9 + *        +                   +                                               +     |
|     | .        |                   |                                               |     |
|   8 + .        +                   +                                               +     |
|     | *.       |                   |                                               | --- |
|   7 + *.       +                   +                                               +     |
|     | **.      |                   |                                               |     |
|   6 + ****.    +                   +                                               +  5  |
|     | *.       |                   |                                               |     |
|   5 + ****.    +                   +                                               +     |
|     | ****     |                   |                                               | --- |
|   4 + *******. +                   +                                               +     |
|     | *.       |                   |                                               |     |
|   3 + ******   +                   +                                               +     |
|     | ****     | Judge 12          |                                               |  4  |
|   2 + ***.     +                   +                                               +     |
|     | *****.   | Judge 9   Judge 7 |                                               |     |
|   1 + ***.     + Judge 2           +                                               +     |
|     | *******  | Judge 3           | (R)Vocabulary & Grammar                       |     |
*   0 * *******. * Judge 11  Judge 6 * (R)Interactive Communication  (R)Pronunciation * --- *
|     | ****.    | Judge 5   Judge 8 |                                               |     |
|  -1 + ****.    + Judge 1   Judge 10+ (I)Holistic Score                             +     |
|     | ***      |                   |                                               |     |
|  -2 + ****.    +                   +                                               +  3  |
|     | ***      |                   |                                               |     |
|  -3 + *****.   +                   +                                               +     |
|     | *****    | Judge 4           |                                               | --- |
|  -4 + **       +                   +                                               +     |
|     | *.       |                   |                                               |     |
|  -5 + *.       +                   +                                               +     |
|     | .        |                   |                                               |     |
|  -6 + .        +                   +                                               +  2  |
|     | .        |                   |                                               |     |
|  -7 + *        +                   +                                               +     |
|     | *.       |                   |                                               |     |
|  -8 +          +                   +                                               + (1) |
|-----+----------+-------------------+-----------------------------------------------+-----|
|Measr| * = 3    |-Judges            |-Items                                         |RATIN|
+------------------------------------------------------------------------------------------+
```

Figure 3 is the Facets ruler from the 2018 BEST 1. The leftmost column shows the Rasch measures, which are linear and true interval, while the rightmost column shows the converted rating scale points, which may or may not be linear. The ruler shows a wide dispersion in student performance, resembling a flattened bell curve.

Through this analysis, the GEAC has a means to evaluate the BEST. For example, the 2018 BEST 1 ruler reveals that analytic ratings were slightly more difficult than the holistic score, meaning it was slightly more difficult for students to earn a high score on Vocabulary & Grammar than on the single holistic rating they received from their interlocutor. The Judges clearly vary in severity more than desired. In fact, the distance between Judge 12 and Judge 4 is 6 Rasch units, a difference of about 1.5 rating scale points on average. The area encompassing 4 points is wider than the areas encompassing 3 points and 5 points (3.5, 4, and 4.5 in the original scale, see Table 6). This means that raters, on average, needed to see a greater change in student performance to award a score of 4.5 than they did to move from 3.5 to 4. Although they are quite close, the rating scale points in practice are not linear.

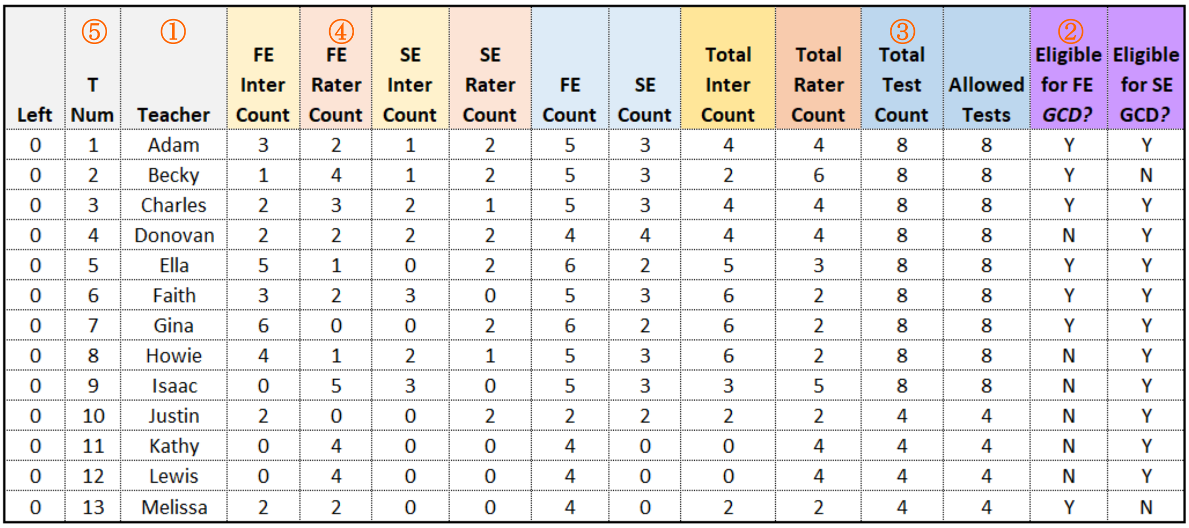
Lesson 2: Creating a scheduling and data entry database

Toward facilitating these scheduling guidelines, it was imperative to build a database where various scheduling permutations could be attempted until all the rules were successfully applied. Starting from the 2016 BEST 2 and 4, the BECC began using a new automated Excel-based scheduling system that utilized several formula-based checks to ensure guideline cooperation. Note that in all subsequent figures, all displayed teacher and student names have been fabricated for anonymity.

As shown in the model plan in Figure 4, the system is set up with teachers ① and the GCD grade levels ② they teach listed in the judging plan section of the scheduling system. Each teacher is also given an ideal test count ③ number based on the average amount of test sessions required in the exam period along with rater or interlocutor designations ④. Finally, each teacher is assigned an index number ⑤ that will later be used to streamline the scheduling process. Counts of FE, SE, rater, interlocutor, total, and remaining test slots are all updated automatically.

Figure 4

2019 BEST 2 and 4 judging plan



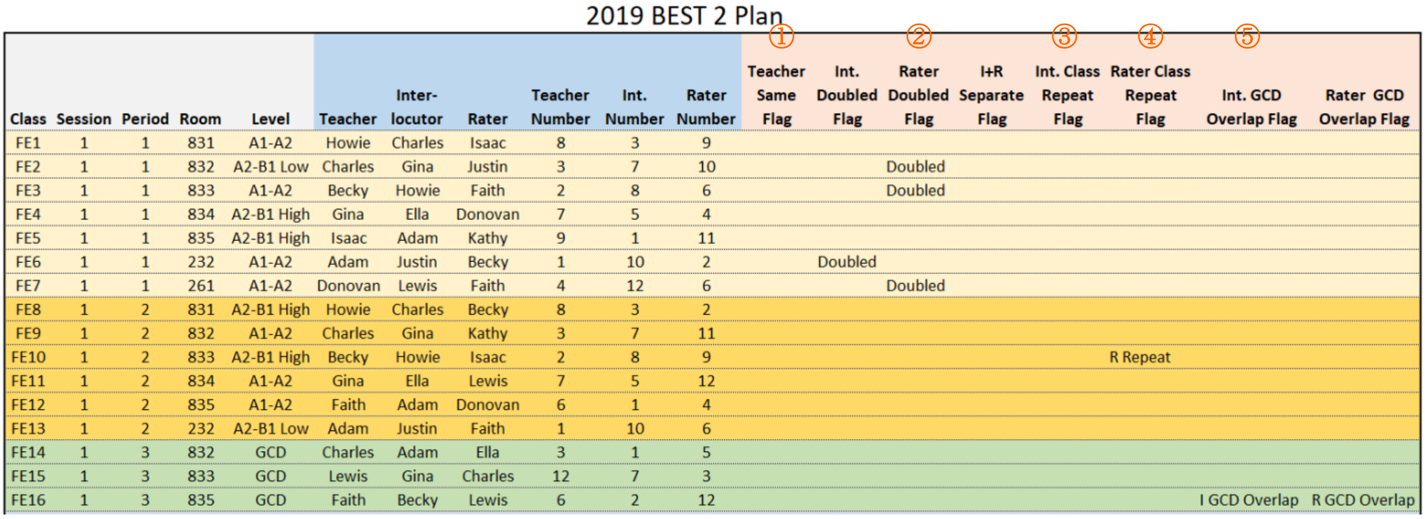
	⑤	①		④							③		②	
Left	T Num	Teacher	FE Inter Count	FE Rater Count	SE Inter Count	SE Rater Count	FE Count	SE Count	Total Inter Count	Total Rater Count	Total Test Count	Allowed Tests	Eligible for FE *GCD?*	Eligible for SE *GCD?*
0	1	Adam	3	2	1	2	5	3	4	4	8	8	Y	Y
0	2	Becky	1	4	1	2	5	3	2	6	8	8	Y	N
0	3	Charles	2	3	2	1	5	3	4	4	8	8	Y	Y
0	4	Donovan	2	2	2	2	4	4	4	4	8	8	N	Y
0	5	Ella	5	1	0	2	6	2	5	3	8	8	Y	Y
0	6	Faith	3	2	3	0	5	3	6	2	8	8	Y	Y
0	7	Gina	6	0	0	2	6	2	6	2	8	8	Y	Y
0	8	Howie	4	1	2	1	5	3	6	2	8	8	N	Y
0	9	Isaac	0	5	3	0	5	3	3	5	8	8	N	Y
0	10	Justin	2	0	0	2	2	2	2	2	4	4	N	Y
0	11	Kathy	0	4	0	0	4	0	0	4	4	4	N	Y
0	12	Lewis	0	4	0	0	4	0	0	4	4	4	N	Y
0	13	Melissa	2	2	0	0	4	0	2	2	4	4	Y	N

BEST raters and interlocutors are scheduled in the Judging Plan section of the system, a portion of which is shown in Figure 5. Each color group of rows (three of which are shown in Figure 5) contains a unique date and period testing block. Regular class teachers, interlocutors, and raters are entered via number, which is linked via formula to the judging plan section of the tab. To the right, several flag columns will populate with warnings if the following guidelines are broken:

- Teacher Same Flag: The classroom teacher has been scheduled for their own class.
- Int. / Rater Doubled Flag: The interlocutor or rater is scheduled twice within the same test session.
- I + R Separate Flag: This combination of interlocutor and rater is already found within this test (among all sessions and dates).
- Int. / Rater Class Repeat Flag: The interlocutor / rater has been scheduled for the same class twice.
- Int. / Rater GCD Overlap Flag: The interlocutor / rater has been scheduled for GCD students whom they potentially teach separately in another course.

Figure 5

2019 BEST 2 schedule portion



2019 BEST 2 Plan

											①		②		③	④	⑤	
Class	Session	Period	Room	Level	Teacher	Inter-locutor	Rater	Teacher Number	Int. Number	Rater Number	Teacher Same Flag	Int. Doubled Flag	Rater Doubled Flag	I+R Separate Flag	Int. Class Repeat Flag	Rater Class Repeat Flag	Int. GCD Overlap Flag	Rater GCD Overlap Flag
FE1	1	1	831	A1-A2	Howie	Charles	Isaac	8	3	9								
FE2	1	1	832	A2-B1 Low	Charles	Gina	Justin	3	7	10			Doubled					
FE3	1	1	833	A1-A2	Becky	Howie	Faith	2	8	6			Doubled					
FE4	1	1	834	A2-B1 High	Gina	Ella	Donovan	7	5	4								
FE5	1	1	835	A2-B1 High	Isaac	Adam	Kathy	9	1	11								
FE6	1	1	232	A1-A2	Adam	Justin	Becky	1	10	2		Doubled						
FE7	1	1	261	A1-A2	Donovan	Lewis	Faith	4	12	6			Doubled					
FE8	1	2	831	A2-B1 High	Howie	Charles	Becky	8	3	2								
FE9	1	2	832	A1-A2	Charles	Gina	Kathy	3	7	11								
FE10	1	2	833	A2-B1 High	Becky	Howie	Isaac	2	8	9						R Repeat		
FE11	1	2	834	A1-A2	Gina	Ella	Lewis	7	5	12								
FE12	1	2	835	A1-A2	Faith	Adam	Donovan	6	1	4								
FE13	1	2	232	A2-B1 Low	Adam	Justin	Faith	1	10	6								
FE14	1	3	832	GCD	Charles	Adam	Ella	3	1	5								
FE15	1	3	833	GCD	Lewis	Gina	Charles	12	7	3								
FE16	1	3	835	GCD	Faith	Becky	Lewis	6	2	12							I GCD Overlap	R GCD Overlap

In an ideal schedule, all rules will have been accommodated and thus all flag columns would be empty. However, variables such as the number of available teachers, the number of simultaneous classes, and teacher eligibility are in flux year by year and may make it impossible to create a schedule that follows all the guidelines. When this happens, priority is placed on minimizing rule-breaking flags over roles (interlocutor only / rater only / hybrid interlocutor + rater) as the now-inherent data linkages make it exceedingly unlikely for Facets to break the data into subsets even when these roles are only partially realized.

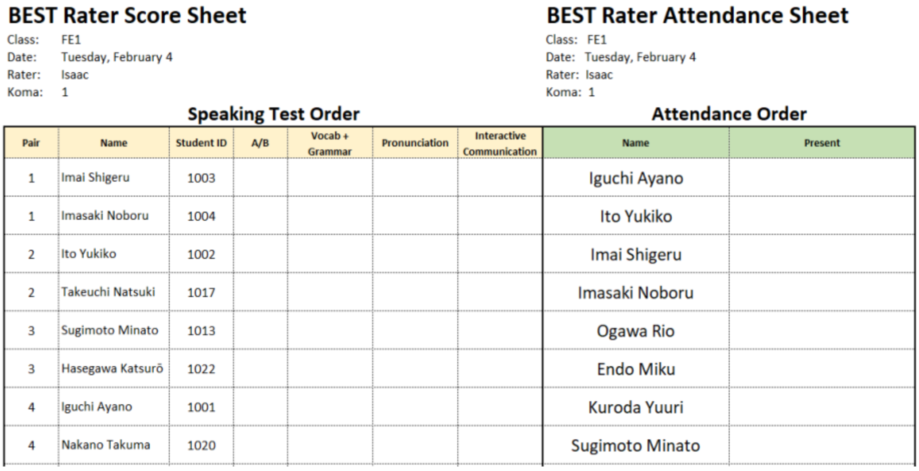
The 2019 BEST 2 and 4 plan as shown in Figures 4 and 5 above, contained 30 FE and 14 SE BEST sessions. In Figure 4, each of the 13 teachers was assigned to four or eight test sessions. Three teachers worked only as raters within FE, while two teachers were only interlocutors, and three other teachers had only a single test session in one role with the remaining sessions as the opposite role. Conversely, five teachers had an even or roughly even number of FE rater and interlocutor sessions. In most cases, the roles were reversed for SE tests. As seen by the blank I+R Separate flag column, no rater-interlocutor pairing was repeated across the exam. However, there were some scheduling shortcomings. In Figure 5, an excess of FE classes resulted in not enough teachers being available to fill all slots (hence the ‘doubled’ flag arising), requiring rating by video camera. Likewise, one rater needed to rate the same class two times (R repeat), and one GCD class saw both a rater and an interlocutor who taught these students in other classes (I / R GCD Overlap). Despite these, the test was successfully facilitated, and the Facets data set was connected.

With this scheduling system in place, attention was turned to the user input system. Through the 2016 BEST 1 and 3, Google Sheets was used to facilitate the BEST score input system. Beginning of term rosters were copied to a single Google Sheet for FE and SE courses, and teachers were required to input the students’ test date, pairing number, and scores. Although the system was adequate, feedback from BECC teachers indicated several aspects of dissatisfaction. First, the class rosters were based on the beginning-of-year streaming document and were often out-of-date due to withdrawals, leading to a multitude of inquiries regarding absences from the judges to the class teacher. Second, as the rosters were listed in student ID order while the actual testing session was in randomized order and spread between two dates, teachers found it taxing and error-prone to find students and transfer the correct testing information and scores. Finally, the system contained no method for the class teacher to create randomized testing and attendance rosters. Rather, these needed to be typed into a separate document, increasing the necessary preparation time and introducing the possibility of double listing or omitting a student.

As a result, from the 2016 BEST 2 and 4, in conjunction with the scheduling system, a new BEST Excel roster creation and data reporting system was created. This system fixed these issues by utilizing three roster tabs in addition to housing the test scheduling system. The first tab houses the beginning of year streaming list, which serves to connect student names, classes, and ID numbers to further roster tabs. Second, as shown in Figure 6 below, each class has its own roster tab for each testing date, where student names are entered in their testing order, and columns for the rater's three scores are provided.

Figure 6

BEST roster 2: Class roster tab



BEST Rater Score Sheet

Class: FE1
Date: Tuesday, February 4
Rater: Isaac
Koma: 1

Speaking Test Order

Pair	Name	Student ID	A/B	Vocab + Grammar	Pronunciation	Interactive Communication
1	Imai Shigeru	1003				
1	Imasaki Noboru	1004				
2	Ito Yukiko	1002				
2	Takeuchi Natsuki	1017				
3	Sugimoto Minato	1013				
3	Hasegawa Katsurō	1022				
4	Iguchi Ayano	1001				
4	Nakano Takuma	1020				

BEST Rater Attendance Sheet

Class: FE1
Date: Tuesday, February 4
Rater: Isaac
Koma: 1

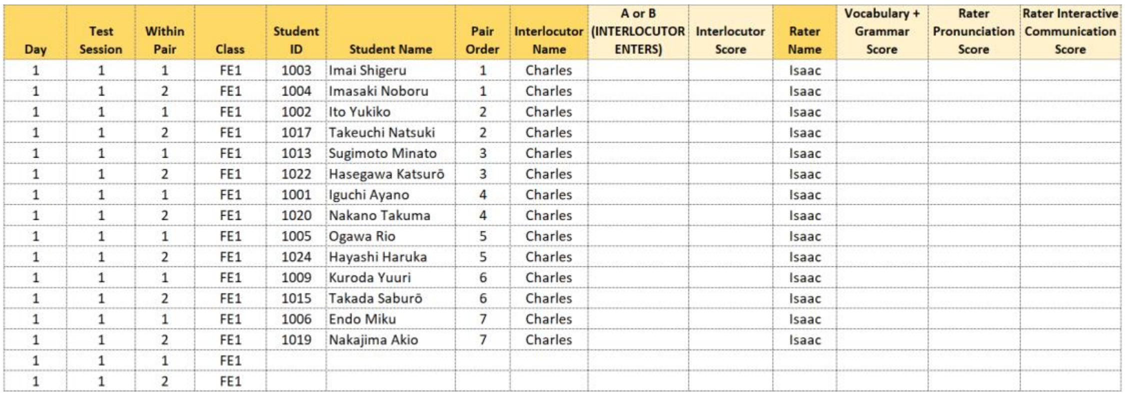
Attendance Order

Name	Present
Iguchi Ayano	
Ito Yukiko	
Imai Shigeru	
Imasaki Noboru	
Ogawa Rio	
Endo Miku	
Kuroda Yuuri	
Sugimoto Minato	

The student IDs are pulled via formula from the streaming list, with an error notification displaying if a student name is misspelled. Utilizing a student ID ranking formula, these names are replicated to the right but in student ID (attendance) order. This roster doubles as the rater scoresheet and attendance checklist, and both rosters are printed and provided to the rater after input is complete. These tabs are connected to the judge scheduling system tab via a matching class and date index, so rater names are automatically listed. Finally, the test rosters for all classes are consolidated into a final score input tab. All columns except for student scores are populated via formulas aligning with a class, date, and order index to pull data from the class roster tabs and BEST scheduling tab (Figure 7).

Figure 7

BEST roster 3: Score input tab



Day	Test Session	Within Pair	Class	Student ID	Student Name	Pair Order	Interlocutor Name	A or B (INTERLOCUTOR ENTERS)	Interlocutor Score	Rater Name	Vocabulary + Grammar Score	Rater Pronunciation Score	Rater Interactive Communication Score
1	1	1	FE1	1003	Imai Shigeru	1	Charles			Isaac			
1	1	2	FE1	1004	Imasaki Noboru	1	Charles			Isaac			
1	1	1	FE1	1002	Ito Yukiko	2	Charles			Isaac			
1	1	2	FE1	1017	Takeuchi Natsuki	2	Charles			Isaac			
1	1	1	FE1	1013	Sugimoto Minato	3	Charles			Isaac			
1	1	2	FE1	1022	Hasegawa Katsurō	3	Charles			Isaac			
1	1	1	FE1	1001	Iguchi Ayano	4	Charles			Isaac			
1	1	2	FE1	1020	Nakano Takuma	4	Charles			Isaac			
1	1	1	FE1	1005	Ogawa Rio	5	Charles			Isaac			
1	1	2	FE1	1024	Hayashi Haruka	5	Charles			Isaac			
1	1	1	FE1	1009	Kuroda Yuuri	6	Charles			Isaac			
1	1	2	FE1	1015	Takada Saburō	6	Charles			Isaac			
1	1	1	FE1	1006	Endo Miku	7	Charles			Isaac			
1	1	2	FE1	1019	Nakajima Akio	7	Charles			Isaac			
1	1	1	FE1										
1	1	2	FE1										

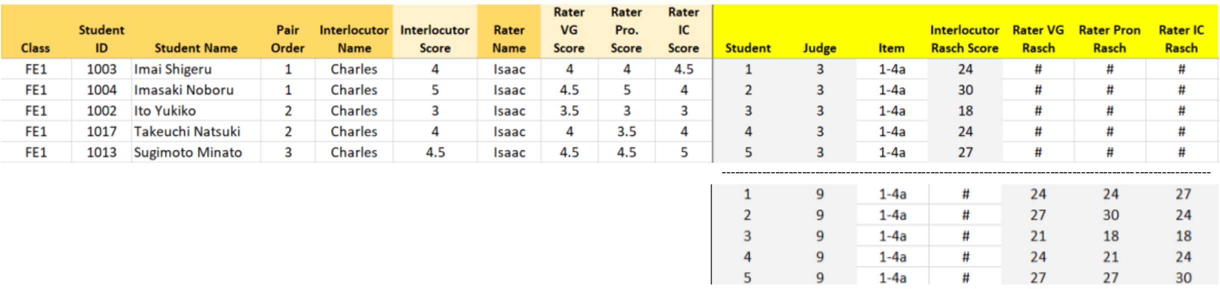
Sixteen rows (half the maximum class size) are stacked sequentially for each class and day, with rows without students intentionally kept blank: thus, a designated row for each student from Day 1, Student 1-A to Day 2, Student 8-B is assigned and only filled in if such student designation exists in each class roster tab. As a result, students are sorted correctly into their actual testing date and order rather than by student ID when teachers open the document to input the scores, easing the reporting process and limiting the potential for data entry errors. Through this automation, the required teacher interaction with the document is minimized, negating the risk of typing errors or doubling or missing students.

Lesson 3: Generating the fairest Rasch fair scores

The next hurdle centered on how to best process fair scores from the Rasch analysis. The scoring input system converts the raw scores and judges into Facets-compatible data lines, as demonstrated in Figure 8. As introduced previously in Table 3, because two separate judges provide one combined set of scores, each judge's score line is recorded in the Rasch input file on a separate line, with the non-applying set of scores listed as "missing" data. In the below example, the interlocutor 'Charles' and rater 'Isaac' are converted by the system to judge numbers 3 and 9 in their respective data lines.

Figure 8

BEST Rasch score converter example



Class	Student ID	Student Name	Pair Order	Interlocutor Name	Interlocutor Score	Rater Name	Rater VG Score	Rater Pro. Score	Rater IC Score	Student	Judge	Item	Interlocutor Rasch Score	Rater VG Rasch	Rater Pron Rasch	Rater IC Rasch
FE1	1003	Imai Shigeru	1	Charles	4	Isaac	4	4	4.5	1	3	1-4a	24	#	#	#
FE1	1004	Imasaki Noboru	1	Charles	5	Isaac	4.5	5	4	2	3	1-4a	30	#	#	#
FE1	1002	Ito Yukiko	2	Charles	3	Isaac	3.5	3	3	3	3	1-4a	18	#	#	#
FE1	1017	Takeuchi Natsuki	2	Charles	4	Isaac	4	3.5	4	4	3	1-4a	24	#	#	#
FE1	1013	Sugimoto Minato	3	Charles	4.5	Isaac	4.5	4.5	5	5	3	1-4a	27	#	#	#
										1	9	1-4a	#	24	24	27
										2	9	1-4a	#	27	30	24
										3	9	1-4a	#	21	18	18
										4	9	1-4a	#	24	21	24
										5	9	1-4a	#	27	27	30

Note: VG = Vocabulary and Grammar; Pron. = Pronunciation; IC = Interactive Communication

The BEST uses seven scoring levels for all categories, consisting of the integers 1-5 and the half marks 3.5 and 4.5 (see Table 1). A complication arose, however, due to the Facets rating scale being unable to process half marks or decimal points, so beginning with the 2016 BEST 2 and 4, BEST scores were converted into sequential integers for the Facets rating scale. A converted score for a mark of 1, indicating a student's refusal to take the exam (see the Appendix), was not assigned due to this score never having been awarded in practice. This left six scoring categories, and accordingly, the following *R6* Facets rating scale was used (Table 6 and Figure 9).

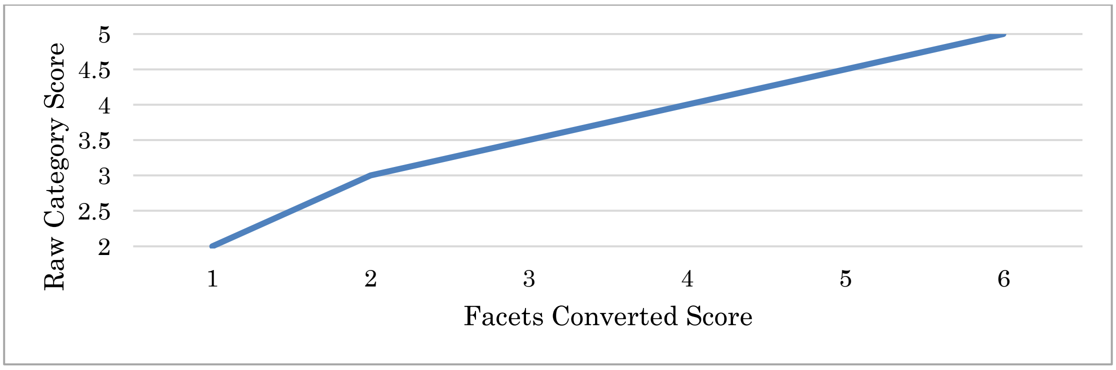
Table 6

2016-18 BEST rating scale (raw and Facets converted scores)

BEST Raw Category Score	Facets Converted Observed Score (Rating Scale)
1	-
2	1
3	2
3.5	3
4	4
4.5	5
5	6

Figure 9

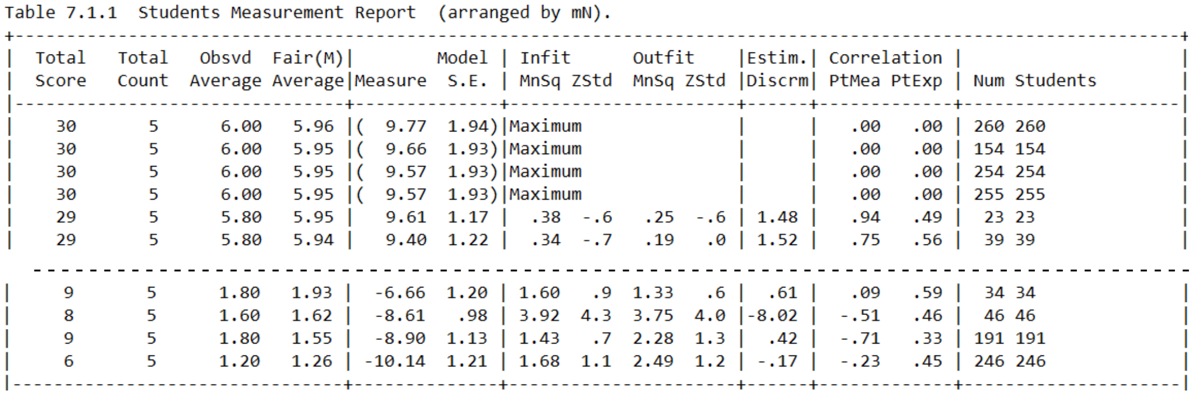
2016-18 BEST rating scale (raw and Facets converted scores)



A further complicating factor was the BEST's implicit requirement that the final scores equate to the actual awarded 15% course grade. Prior to instituting the MFRM, this grade was a simple sum of the four raw scores (with the holistic score double weighted), producing a score out of 25, then multiplied by 0.6 to make a final score out of 15. However, the converted Facets observed score rating scale, with a maximum of 6 points per category, was not a linear conversion from the original raw scores, so it was not possible simply to reconvert the Facets fair scores back to real averages by multiplying by 2.5 to achieve a score out of 15. Therefore, rather than the MFRM fair scores, the GEAC utilized Rasch logit measures as the ultimate grade and converted them via UMEAN (Linacre, 2022b) to a scale of 6 to 15, with the minimum score 6 being equivalent to the lowest possible observed BEST score average of 2 out of 5 for each scoring category (a raw 10 out of 25, multiplied by 0.6 to arrive at 6). Calculating the UMEAN requires determining the mean of all measures and the points per logit. To calculate the points per logit, the scoring range (nine) was divided by the logit range, or the absolute value of the combined top and bottom student measures. The mean of all measures is comprised of the absolute scoring range (15) minus the product of the points per logit and the lowest measure to receive a maximum score (Linacre, 2022b). An example, taken from the 2017 BEST 4, is shown in Figure 10 and Table 7.

Figure 10

2017 BEST 4 top and bottom measures



```
Table 7.1.1  Students Measurement Report  (arranged by mN).
+----------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit       Outfit     |Estim.| Correlation |                   |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd   MnSq ZStd  |Discrm| PtMea PtExp | Num Students      |
|--------------------------------+--------------+------------------------+------+-------------+-------------------|
|    30      5      6.00    5.96 |(  9.77  1.94)|Maximum                 |      |   .00   .00 | 260 260           |
|    30      5      6.00    5.95 |(  9.66  1.93)|Maximum                 |      |   .00   .00 | 154 154           |
|    30      5      6.00    5.95 |(  9.57  1.93)|Maximum                 |      |   .00   .00 | 254 254           |
|    30      5      6.00    5.95 |(  9.57  1.93)|Maximum                 |      |   .00   .00 | 255 255           |
|    29      5      5.80    5.95 |   9.61  1.17 |  .38  -.6    .25  -.6  | 1.48 |   .94   .49 |  23 23            |
|    29      5      5.80    5.94 |   9.40  1.22 |  .34  -.7    .19   .0  | 1.52 |   .75   .56 |  39 39            |

   ---------------------------------------------------------------------------------------------------------------
|     9      5      1.80    1.93 |  -6.66  1.20 | 1.60   .9   1.33   .6  |  .61 |   .09   .59 |  34 34            |
|     8      5      1.60    1.62 |  -8.61   .98 | 3.92  4.3   3.75  4.0  |-8.02 |  -.51   .46 |  46 46            |
|     9      5      1.80    1.55 |  -8.90  1.13 | 1.43   .7   2.28  1.3  |  .42 |  -.71   .33 | 191 191           |
|     6      5      1.20    1.26 | -10.14  1.21 | 1.68  1.1   2.49  1.2  | -.17 |  -.23   .45 | 246 246           |
|--------------------------------+--------------+------------------------+------+-------------+-------------------|
```

Adding the absolute value of the top (A) and bottom (B) measures as shown in Figure 10 resulted in a logit range (C) of 19.91, which was multiplied by 9 (D, the actual score range of 15 minus 6) to result in a .452 points per logit calculation (E). Multiplying the lowest full score measure (F) by the points per logit to form G and subtracting that value from the absolute scoring range (H, or 15), resulted in a mean of all measures of 10.674. Thus, the final UMEAN code line input into the Facets input file is *UMEAN=10.674,.452,2*. In Figure 11, the UMEAN adjustment now provided a top and bottom measure range of roughly 15 to 6, which was utilized as the student exam grade range.

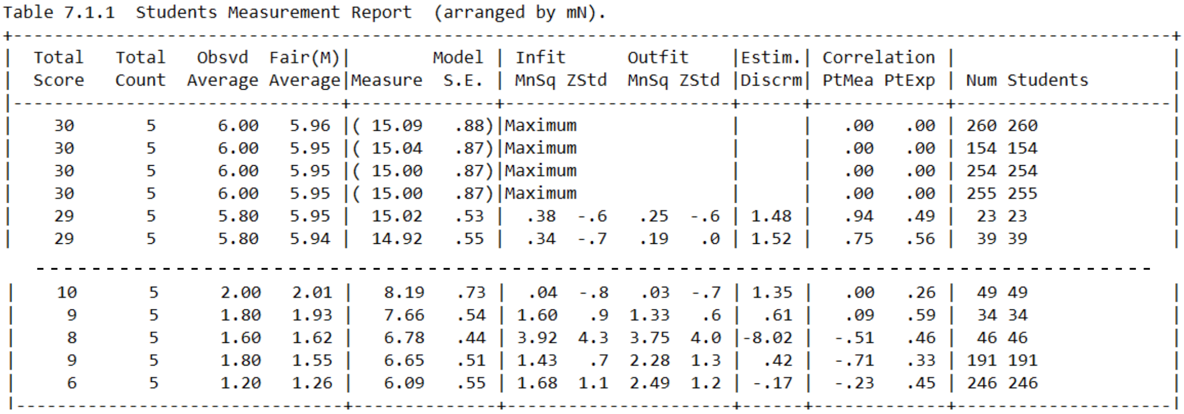
Table 7

UMEAN scoring calculation example (2017 BEST 4 data)

Points per Logit Calculation:	High Logit Measure:	Low Logit Measure:	Logit Range [|A+B|]:	Max–Min Actual Score [15-6]:	Points per Logit [C*D]:
	9.77 (A)	-10.14 (B)	19.91 (C)	9 (D)	.452 (E)
Mean of all Measures Calculation:	Lowest Full Score Measure:	Lowest Full Score Points[E*F]:	Absolute Scoring Range:	Mean of all Measures[H-G]:	
	9.57 (F)	4.32 (G)	15 (H)	10.674 (I)	

Figure 11

UMEAN adjusted 2017 BEST 4 top and bottom measures

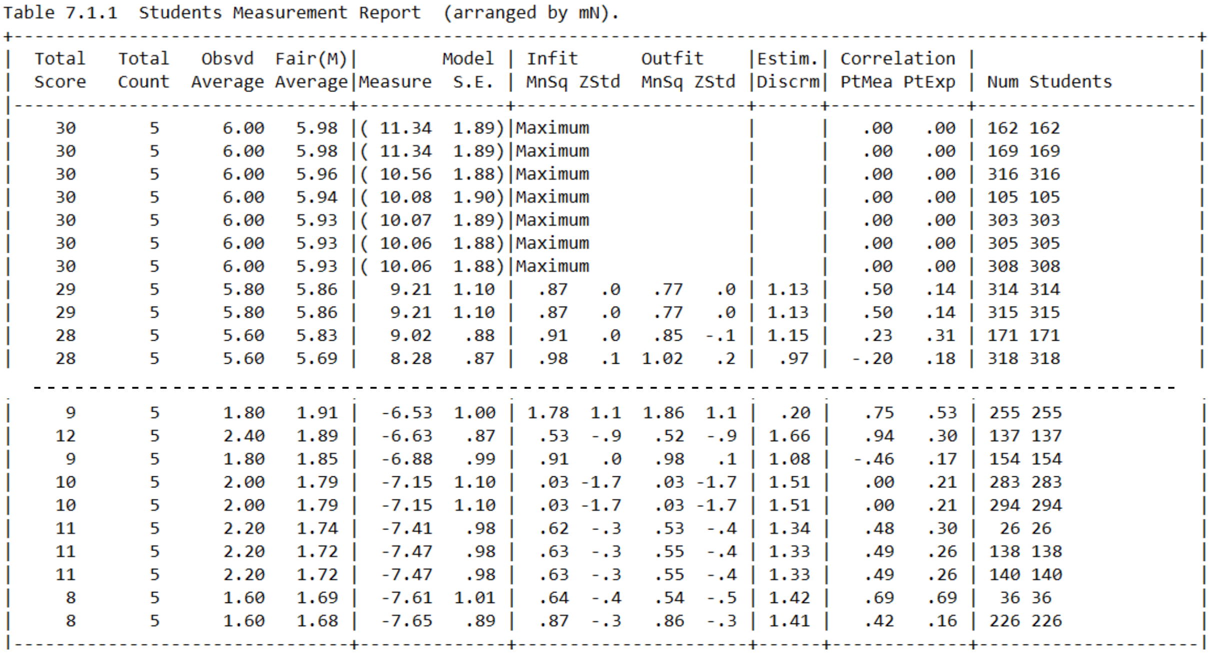


```
Table 7.1.1  Students Measurement Report  (arranged by mN).
+-----------------------------------------------------------------------------------------------------+
| Total   Total   Obsvd  Fair(M)|        Model | Infit       Outfit     |Estim.| Correlation |                      |
| Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Num Students         |
|-------------------------------+--------------+----------------------+------+-------------+----------------------|
|   30      5      6.00   5.96 |( 15.09   .88)|Maximum               |      |   .00   .00 | 260 260              |
|   30      5      6.00   5.95 |( 15.04   .87)|Maximum               |      |   .00   .00 | 154 154              |
|   30      5      6.00   5.95 |( 15.00   .87)|Maximum               |      |   .00   .00 | 254 254              |
|   30      5      6.00   5.95 |( 15.00   .87)|Maximum               |      |   .00   .00 | 255 255              |
|   29      5      5.80   5.95 |  15.02   .53 |  .38  -.6   .25  -.6 | 1.48 |   .94   .49 |  23 23               |
|   29      5      5.80   5.94 |  14.92   .55 |  .34  -.7   .19   .0 | 1.52 |   .75   .56 |  39 39               |

   --------------------------------------------------------------------------------------------

|   10      5      2.00   2.01 |   8.19   .73 |  .04  -.8   .03  -.7 | 1.35 |   .00   .26 |  49 49               |
|    9      5      1.80   1.93 |   7.66   .54 | 1.60   .9  1.33   .6 |  .61 |   .09   .59 |  34 34               |
|    8      5      1.60   1.62 |   6.78   .44 | 3.92  4.3  3.75  4.0 |-8.02 |  -.51   .46 |  46 46               |
|    9      5      1.80   1.55 |   6.65   .51 | 1.43   .7  2.28  1.3 |  .42 |  -.71   .33 | 191 191              |
|    6      5      1.20   1.26 |   6.09   .55 | 1.68  1.1  2.49  1.2 | -.17 |  -.23   .45 | 246 246              |
|-------------------------------+--------------+----------------------+------+-------------+----------------------|
```

While this measure-based calculation provided the desired scoring range, two issues that became prevalent in some tests were high-end outliers and the underutilization of the low end of the rating scale, which skewed the bell curve of results. A student awarded a perfect score (possibly due to the test assessing only up to the CEFR B1 level) despite having strict raters could consume the top echelon of the rating scale, resulting in UMEAN converted measures that were nearly universally lower than the initial observed averages regardless of adjustment due to judge leniency and severity. On the other hand, if the low end of the rating scale went underutilized, spreading the expected fair measures between 15 to 6 would tend to stretch students downwards, with minor observed gaps between student scores stretched to larger ones, as Facets used teacher leniency and severity to tier students on the 15 to 6 scale. This was particularly worrisome; in practice, final grades lower than 9 were originally quite rare, as judges only sparingly gave Facets converted 1-point scores (unconverted 2-point scores). This weakness was borne out of the unfortunate fact that the calculated exam scores slotted directly in as student grades, and thus the final grades needed to mirror the raw scores. In such cases, judgement calls on whether to shorten the rating scale or otherwise modify the UMEAN calculation to arrive at a more ideal scoring curve needed to be made case by case, leading to inconsistent calculations between exams. Furthermore, the amount of time required to hone the calculation and the subjectivity in forcing the converted measures to meet a desired bell curve necessitated a rethinking of the calculation procedures. Taken from the 2018 BEST 1, Figure 12, in conjunction with the ruler in Figure 3 above, demonstrates some of these difficulties.

Figure 12

2018 BEST 1 top and bottom measures



```
Table 7.1.1  Students Measurement Report  (arranged by mN).
+-------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd  Fair(M)|        Model | Infit       Outfit     |Estim.| Correlation |                     |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd   MnSq ZStd  |Discrm| PtMea PtExp | Num Students        |
|--------------------------------+--------------+------------------------+------+-------------+---------------------|
|    30      5      6.00    5.98 |( 11.34  1.89)|Maximum                 |      |   .00   .00 | 162 162             |
|    30      5      6.00    5.98 |( 11.34  1.89)|Maximum                 |      |   .00   .00 | 169 169             |
|    30      5      6.00    5.96 |( 10.56  1.88)|Maximum                 |      |   .00   .00 | 316 316             |
|    30      5      6.00    5.94 |( 10.08  1.90)|Maximum                 |      |   .00   .00 | 105 105             |
|    30      5      6.00    5.93 |( 10.07  1.89)|Maximum                 |      |   .00   .00 | 303 303             |
|    30      5      6.00    5.93 |( 10.06  1.88)|Maximum                 |      |   .00   .00 | 305 305             |
|    30      5      6.00    5.93 |( 10.06  1.88)|Maximum                 |      |   .00   .00 | 308 308             |
|    29      5      5.80    5.86 |   9.21  1.10 |  .87    .0    .77   .0 | 1.13 |   .50   .14 | 314 314             |
|    29      5      5.80    5.86 |   9.21  1.10 |  .87    .0    .77   .0 | 1.13 |   .50   .14 | 315 315             |
|    28      5      5.60    5.83 |   9.02   .88 |  .91    .0    .85  -.1 | 1.15 |   .23   .31 | 171 171             |
|    28      5      5.60    5.69 |   8.28   .87 |  .98    .1   1.02   .2 |  .97 |  -.20   .18 | 318 318             |

   ------------------------------------------------------------------------------------------------------------

|     9      5      1.80    1.91 |  -6.53  1.00 | 1.78   1.1   1.86  1.1 |  .20 |   .75   .53 | 255 255             |
|    12      5      2.40    1.89 |  -6.63   .87 |  .53   -.9    .52  -.9 | 1.66 |   .94   .30 | 137 137             |
|     9      5      1.80    1.85 |  -6.88   .99 |  .91    .0    .98   .1 | 1.08 |  -.46   .17 | 154 154             |
|    10      5      2.00    1.79 |  -7.15  1.10 |  .03  -1.7    .03 -1.7 | 1.51 |   .00   .21 | 283 283             |
|    10      5      2.00    1.79 |  -7.15  1.10 |  .03  -1.7    .03 -1.7 | 1.51 |   .00   .21 | 294 294             |
|    11      5      2.20    1.74 |  -7.41   .98 |  .62   -.3    .53  -.4 | 1.34 |   .48   .30 |  26 26              |
|    11      5      2.20    1.72 |  -7.47   .98 |  .63   -.3    .55  -.4 | 1.33 |   .49   .26 | 138 138             |
|    11      5      2.20    1.72 |  -7.47   .98 |  .63   -.3    .55  -.4 | 1.33 |   .49   .26 | 140 140             |
|     8      5      1.60    1.69 |  -7.61  1.01 |  .64   -.4    .54  -.5 | 1.42 |   .69   .69 |  36 36              |
|     8      5      1.60    1.68 |  -7.65   .89 |  .87   -.3    .86  -.3 | 1.41 |   .42   .16 | 226 226             |
|--------------------------------+--------------+------------------------+------+-------------+---------------------|
```

In the 2018 BEST 1, MFRM calculated a 2.13 logit ability gap between the highest ability student with a perfect observed total score and the first student with a less than perfect score (students 162 and 314, whose measures were 11.34 and 9.21 respectively). In addition, there was nearly a full logit difference between the lowest perfect total observed score (student 308, whose measure was 10.06) and student 314's 9.21. This means that despite their observed scores being nearly identical, student 314 and all students below saw their measures stretched lower on the rating scale to account for this discrepancy. On the other hand, there was a much blurrier picture with the bottom measures. Any student with a Facets-converted total score of less than 10 received a rare Facets-converted 1 mark in one or more scoring categories, yet Figure 12 shows that rater leniency and severity determined these students to have performed at similar measures to those who received at least a raw 3 (Facets-converted 2) mark in each category. Thus, those who performed at the bottom rung among observed scores and would have normally alone made up the converted 6-8 range of the score ladder were mixed in with those who scored in the 9-10 range. Since the UMEAN is set to assign students a measure between 15 to 6, the 6-8 range of the score ladder is expanded. As Table 8 and Figure 13 demonstrate, the calculated mean of all measures value of 10.23 for the 2018 BEST 1 resulted in nearly all scores dropping from their original observed values.

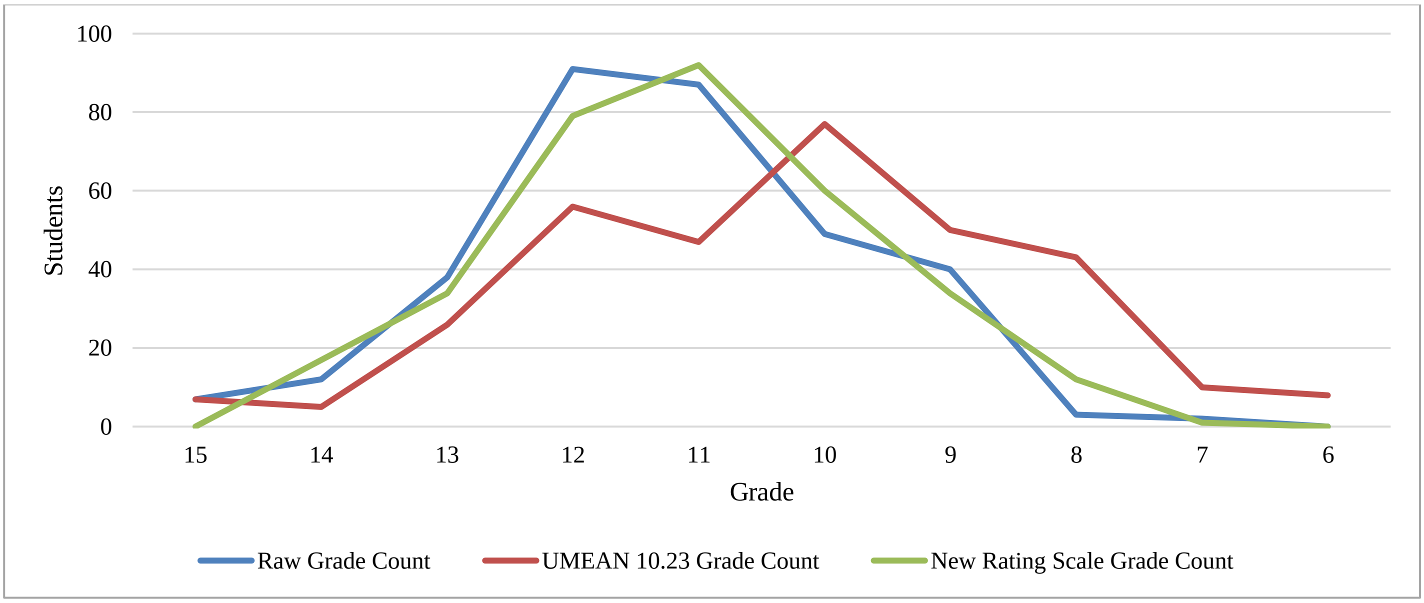
Table 8

2018 BEST 1 UMEAN calculations (n = 329)

Mean of All Measures	Mean Observed vs. New Score Change	Median Observed vs. New Score Change	Max + Observed vs. New Score Change	Max - Observed vs. New Score Change	Score Increases (New vs. Observed)	Score Same (New vs. Observed)	Score Decreases (New vs. Observed)
10.23	-0.93	-0.89	0.61	-2.61	10 (3%)	3 (1%)	316 (96%)
10.5	-0.67	-0.62	0.39	-2.34	35 (11%)	8 (2%)	286 (87%)
11.66	0.45	0.52	1.55	-1.18	268 (81%)	8 (2%)	53 (16%)
New Rating Scale *(See Table 9)*	-0.11	-0.04	0.72	-1.28	147 (45%)	0 (0%)	182 (55%)

Figure 13

2018 BEST bell curves



The UMEAN calculation of 10.23 resulted in all but thirteen students' scores decreasing from their pre-Facets observed average to their fair average, with an average drop of 0.93 points and a median drop of 0.89 points, or nearly 1% of their course grade. Furthermore, the students who gained points were mostly those who already earned a perfect observed score (those at the top of Figure 12), while conversely, as predicted and indicated in Figure 13's rightward shift of the bell curve (see the orange line), students who scored weaker observed scores were dragged further downward. Thus, despite having anchored judges at zero in the MFRM, the measure calculations were serving to reduce student grades, unintentionally defeating the purpose of the fair score calculations.

To counteract this, tweaks were made to the 2018 BEST 1 mean of all measures calculation, resulting in values of 10.5 and later 11.66. The latter value resulted in more favorable student scores for those in the center of the bell curve that more closely aligned with the raw scores, so the 11.66 value was ultimately utilized. However, it was clear that this calculation ambiguity would not be sustainable going forward, and in the summer of 2018, the GEAC began considering alternatives, returning to the rating scale conversion. Rather than using UMEAN-converted measures, it was posited to recalibrate the

Facets converted scores to make them directly linear with the observed scores by multiplying them by six, resulting in an integer-only score range that is divisible by the total points (15) as shown in Table 9 below.

Table 9

2018 onward BEST raw and Rasch converted scores

BEST Observed Category Score	New Converted Facets Observed Score (Rating Scale)
1	-
2	12
3	18
3.5	21
4	24
4.5	27
5	30

While this would circumvent the need for non-integers, which are incompatible with Facets, it had also been assumed until this point that the rating scale value must equal the total number of scoring categories (*R6*), requiring the rating scale to be sequential integers. However, the new plan increased the rating scale to *R30*, which sets Facets up to process 30 scoring points, but only utilized six of them (Table 9), with all other scores reported as *X=0* (or omitted) in the input file. After testing, it was discovered that Facets took no issue with 24 out of 30 scoring categories being blank and unused, calculating statistics only for those reported, a revelation that made the process infinitely easier. The Facets-reported fair scores simply needed to be divided by 2 to be converted into fair grades out of 15 points, with the measures kept for statistical records but not utilized in grading. This made for a consistent scoring system that maintained the original bell curve of the data, and from the 2018 BEST 2 and 4, this new method was adopted.

As shown in Table 8 and Figure 13, when reapplied to the 2018 BEST 1, the new rating scale fair scores matched much more closely with their corresponding observed scores, with a mean and median change of -0.11 and -0.04 points, respectively. The grey line in Figure 13 indicates this moderate scoring shift and keeps nearly the same student ratio at the rightmost end. Furthermore, while the original measure-based calculation saw 96% of scores drop and only 3% increase, the new rating scale method saw a ratio of 55% to 45% respectively, changes much more in line with the expected adjustments due to rater leniency and severity. In addition, the most extreme plusses and minuses in student fair scores were also overall less than any of the three attempted measure-based figures.

Once processed in Facets, the BEST fair scores are extracted from the output file and put back into the BEST score input system, where they are compared against the students' weighted raw scores to determine the volatility of Facets' adjustments for teacher strictness and leniency. Finally, these scores are replicated into a new document for distribution to teachers as well as merged into individual student result cards which teachers distribute in the days following the exam.

Further challenges and conclusion

This process of BEST administrative refinement from 2015 to 2019 of strengthening the rater schedule so that the scores would be a single Facets-compatible data set, building a comprehensive test score and scheduling database, and refining the post-Facets fair score calculation method, helped the BECC in its search for CEFR-aligned exam validity. The GEAC finally had a consistent plan, system, and fair score calculation process.

However, this process also revealed some lingering faults in the application of MFRM to the BEST. First, the outfit mean-square values of some examinees, such as the sample in Figures 10 and 12 above, show values both too low (<0.5, with scores lacking expected variance) or too high (>1.5 or even >2.0, indicating scores with too much variance for the MFRM to show confidence in). These numbers may indicate that the paucity of data points per examinee due to being awarded separate score values by two different judges is resulting in the BEST structure being a poor fit for the MFRM model, or that the judges are being inconsistent in their scoring, necessitating further standardization. In other words, while MFRM does provide fair score calculations, whether they are well-grounded enough to be trusted, particularly from a rater severity standpoint, may require further investigation. Second, as the English Communication curriculum only assesses up to a CEFR B1 level, students above the B1 level are not being accurately assessed by the BEST, causing their measures to be reported as maximum as in Figure 12. To better fit the MFRM so these students' ability levels can be accurately processed with the rest of the cohort, the BEST may need to add a higher scoring category. Finally, one recommendation to enhance the BECC standardization sessions would be to perform a Facets judge bias analysis. Such analysis was conducted in 2016 in the initial MFRM trial period, but performing it regularly would provide the GEAC with further insight into how specific judges are determining and applying their scores.

In 2020 and 2021, the BESTs were cancelled due to COVID-19, and from the 2022 academic year, due to a shift in content facilitation, the FE BESTs are to be replaced by a series of in-class speaking assessments. However, regardless of their long-term continuation, it is hoped that through the lessons learned during this BEST refinement process, the GEAC can continue to improve its exam services and simultaneously help other academic institutions fine-tune their programs toward providing the best possible services to students.

Acknowledgements

I would like to thank the *Shiken* reviewers for their excellent insights and contributions to this paper.

References

Council of Europe (COE). (2018). *Common European Framework of Reference for Languages (CEFR): Learning, teaching, assessment. Companion Volume with New Descriptors*. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Council of Europe (COE). (2001). *Common European Framework of Reference for Languages (CEFR): Learning, teaching, assessment.* Cambridge University Press.

Linacre, J. M. (1997). *MESA Research Note #3: Judging Plans and Facets*. https://www.rasch.org/rn3.htm

Linacre, J. M. (2022a). *Facets computer program for many-facet Rasch measurement, version 3.83.5*. Winsteps.com

Linacre, J. M. (2022b). *Facets Help for Many-Facets Rasch Measurement, Program Manual 3.83.5*, p. 177. https://www.winsteps.com/a/Facets-Manual.pdf

Sugg, R. & Svien, J. (2018). Standardizing Teacher Training for CEFR-based Speaking Assessments. *Bulletin of Hiroshima Bunkyo Women's University*, Volume 53, 45-66.

University of Cambridge ESOL Examinations. (2016). *Cambridge English Key: Key English Test (KET) CEFR Level A2 Handbook for Teachers*. https://www.cambridgeenglish.org/Images/168163-cambridge-english-key-handbook-for-teachers.pdf

University of Cambridge ESOL Examinations. (2016). *Cambridge English Preliminary: Preliminary English Test (PET) CEFR Level B1 Handbook for Teachers*. https://www.cambridgeenglish.org/Images/168150-cambridge-english-preliminary-teachers-handbook.pdf

Appendix

BEST rubrics

CEFR Level	BEST Score	Holistic Interlocutor Rubric (40%)	Analytic Rater Rubrics (60%)		
			Grammar & Vocabulary	*Pronunciation*	*Interactive Communication*
B1 or above	5	Handles communication in **everyday** situations, **despite** hesitation. Constructs **longer** utterances **but** is **not** able to use complex language **except** in **well -rehearsed** utterances.	Shows a **good degree** of **control** of simple grammatical forms. Uses a **range** of appropriate vocabulary when talking about everyday situations.	Pronunciation **is clear and intelligible**, even if a foreign accent is sometimes evident. **Occasional** mispronunciations, but **always the same** words. Student maintains a **smooth rhythm** with **little if any hesitation**.	**Maintains simple exchanges.** Requires no or very little prompting and support. *May use gestures* ***in addition to*** *correct language to help a partner understand.*
A2+	4.5	*Performance shares features of bands 4 and 5.*			
A2	4	Conveys **basic** meaning in **very familiar everyday** situations. Produces utterances which tend to be very short – **words or phrases** – with **frequent hesitation**.	Shows **sufficient** control of simple grammatical forms. Uses appropriate vocabulary to talk about everyday situations.	Pronunciation is **clear enough to be intelligible**, despite a noticeable foreign accent. **Some** mispronunciations occur. Student maintains a **rhythm within memorized sentences**, but with some hesitation **between** sentences.	Maintains simple exchanges, despite **some difficulty**. Requires prompting and support. *May* ***need to use some gestures in lieu of correct language*** *to help a partner understand*
A1+	3.5	*Performance shares features of bands 3 and 4.*			
A1	3	Has **difficulty conveying** basic meaning **even** in very familiar everyday situations. Responses are **limited** to **short phrases or isolated words** with **frequent hesitation and pauses**.	Shows only **limited control** of grammatical forms. Uses a vocabulary of **isolated** words and phrases.	Can be understood with **some effort** by native speakers used to dealing with speakers of this language group. **Many** mispronunciations occur. Student is **monotone** in rhythm, **frequently hesitates** and/or speaks in **broken phrases**.	Has **considerable difficulty** maintaining simple exchanges. Requires additional prompting and support. *May need to* ***rely on gestures to communicate****.*
Pre-A1	2	**Unable to produce the language** to complete the tasks.	Shows **no control** of grammatical forms. Uses **inappropriate** vocabulary or **mostly** L1**.**	Pronunciation is **mostly unintelligible** and / or **impedes communication.**	Unable to ask or respond to most questions.
Pre-A1	1	*Does not attempt the task.*			