SHIKEN

A Journal of Language Testing and Evaluation in Japan

Volume 26 • Number 1 • June 2022

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

Contents

JALT

Testing and Evaluation SIG

ISSN 1881-5537

Shiken: A Journal of Language Testing and Evaluation in Japan

Volume 26 No. 1
June 2022

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

Modeling vocabulary size using many-faceted Rasch measurement

Trevor Holster[1] and J. W. Lake[2]
trevholster@gmail.com
1. Fukuoka University, Fukuoka
2. Fukuoka Jogakuin University, Fukuoka
https://doi.org/10.37546/JALTSIG.TEVAL26.1-1

Abstract

Research into second-language vocabulary size has suffered from inattention to psychometric issues, with ordinal-level raw scores often analyzed as if they represented ratio-level measurement. Additionally, contextual effects have been largely ignored, leading to concern over the interpretation of research findings. This study used many-faceted Rasch measurement to analyze vocabulary data from 1,872 Japanese university students. A test of word synonymy was linked to the *Vocabulary Size Test*, and the contextual variables of item position and time of administration analyzed as measurement facets. Major findings were that data-model fit was sufficient to allow local linking of different item types and contextual variables, allowing meaningful comparison of results and score gains on a scale of vocabulary size, and that item placement within a test form had a substantive effect on item difficulty.

Keywords: Vocabulary size, many-faceted Rasch measurement, test linking, guessing correction

Read (2000) provided a detailed introduction to the nature of vocabulary knowledge, a complex construct that extends beyond simply knowing dictionary definitions. This paper therefore does not attempt to address vocabulary knowledge in its entirety, but is limited to the construct of vocabulary size as operationalized by the *Vocabulary Size Test* (VST) (Beglar, 2010; Nation & Beglar, 2007). Vocabulary size, as described by Chapelle (1994), refers to the number of content words known within a particular context of use, following Dollerup et al.'s (1989) interactionalist view that our comprehension of vocabulary will vary depending on the context in which it is encountered. Chapelle (1998) noted neglect of issues of validity in second language (L2) vocabulary assessment research, a concern that was belatedly acknowledged by Schmitt et al. (2020) over two decades later. The estimation of vocabulary size is one such area of concern.

Intuitively, estimates of vocabulary size should be invariant between repeated test administrations, but this invariance will not hold between raw percentage scores from vocabulary test forms sampling from different frequency ranges. This is because of the unavoidable presence of idiosyncratic words whose difficulty level does not align with their frequency within the target corpus, an effect seen quite dramatically in Beglar's (2010) results. One cause of such idiosyncratic items would be the inclusion of cognates between the students' L1 and L2 in a test, an issue that Read (1988) warned was a threat to the interpretation of test scores if students from different language backgrounds are tested together due to differential item functioning (DIF) of words that are cognates with the L1 of one group but not of other groups. DIF concerns also arise over differential patterns of language exposure or study between different subgroups of students sharing the same L1. Such an effect was reported by Santellices and Wilson (2010), where the different language backgrounds of Black and White American students resulted in DIF on SAT language questions. This DIF is inevitable whenever students from different language backgrounds are tested together so cannot be resolved by changing the corpus used to estimate word frequency. Researchers investigating the relationship between word frequency and test item difficulty must therefore recognize idiosyncratic knowledge as an inescapable feature of language rather than something that can be addressed through corpus sampling design.

The consequence of idiosyncratic knowledge is that vocabulary size estimates will vary between test forms sampling different frequency ranges. To illustrate the problem, if a large number of students were tested on a 5K VST form, with 10 items from each of the first 1000-word bands, students with scores of 25 out of 50 would have an estimated vocabulary size of 2,500 words, but some of those students would also know some lower frequency words. Thus, the vocabulary size estimate of 2,500 words underestimates their vocabulary size, which would be expected to increase if a 14K test was administered, and increase again if a 20K form were administered. Beglar (2010), for instance, administered the lowest group in his study a 40-item 4K form and the middle level group an 80-item 8K version so the vocabulary size estimates of those groups would have been underestimated relative to students taking 14K or 20 K test forms. In principle, any test of a practical length will always underestimate vocabulary size due to the idiosyncratic nature of vocabulary knowledge, so the theorized invariance of vocabulary size cannot be expected to be observed in practice.

Guessing effects

The underestimation due to idiosyncratic knowledge is unrelated to whether random or informed guessing is present or absent, making linking of scores between different versions of the VST necessary if they are to be compared. However, the linking of test forms does require consideration of the effects of random guessing because the 4-option selected response (S.R.) format of the VST means that random guessing alone would give an expected average score of 35 on the original 14K form. This corresponds to a vocabulary size estimate of 3,500 words if Nation's (2012) advice to multiply raw scores by a scaling constant of 100 to obtain a vocabulary size estimate. Nation emphasized that an "I don't know" option was not included in the VST because "the learners should make informed guesses" (2012, p. 4), advice that renders invalid the protocol of estimating vocabulary size by use of a simple scaling constant and has other important implications for the construct definition of vocabulary size.

As Holster and Lake (2016) discussed, guessing correction is a well-established procedure in interpreting scores from multiple-choice tests (Frary, 1988, for example). An important reason for advising students to guess unknown test items is that many students may be confused by technical explanations of guessing strategies, favoring students who adopt optimal strategies over those who do not (Budescu & Bar-Hillel, 1993). Nation's (2012) use of the term "informed guesses" reflects that knowledge is not a simple dichotomy between complete knowledge and zero knowledge. Human knowledge of anything is incomplete, so responses to test questions always reflect partial knowledge, with the probability of success increasing with a candidate's level of partial knowledge. Further, as Thissen et al. (1989) pointed out, distractors are an integral part of a test item, so S.R. vocabulary test scores represent knowledge of the stem, key, and distractors; such test items are not intended to test knowledge of a single target word. In a 4-option S.R. format, eliminating one distractor when the key is unknown increases the probability of guessing the correct response from 25% to 33%, eliminating two distractors increases it to 50%, and eliminating three distractors results in a 100% probability of success. Distractor elimination is thus a construct-relevant display of knowledge and a correct response cannot be assumed to represent knowledge of the item key. Rather than confusing students with technical explanations about when it is appropriate or inappropriate to employ informed guessing, advice which test-wise students are likely to ignore anyway, it is therefore preferable to just instruct them to guess randomly from any response options that they cannot eliminate. The use of an S.R. format coupled with Nation's (2012) endorsement of informed guessing thus has two important consequences for the validity of the VST: i) guessing correction is required to convert raw scores to vocabulary size estimates; ii) the construct is inherently limited to an estimate of how many words a student understands, not whether they understand any specific word included in the test.

Measurement invariance and test linking

The linking and rescaling of different test forms to a reference form requires measurement invariance, meaning that relative person ability is unaffected by the sample of test items used and relative item difficulty is unaffected by the sample of persons tested (Engelhard, 2013). The Rasch measurement model (Rasch, 1960; Wright & Stone, 1979) achieves this invariance through the conversion of raw percentage scores to log-odds unit, or *logits.* This logit conversion is required because raw percentage scores from different test forms or scoring protocols do not provide invariant measurement. Measurement invariance also makes Rasch generated logits useful for measuring the effect size of learning gains calculated through the subtraction of pre-test scores from post-test scores. These subtractive comparisons require an *interval level* measurement scale, following Stevens' (1946) hierarchy of measurement scales, a property of Rasch logit measures but not of raw percentage scores.

Crucially, although Beglar (2010) used Rasch analysis in his validation study, the construct of vocabulary size was defined in terms of raw scores, giving the practical advantage that classroom teachers can administer, score, and interpret the VST without needing any expertise in psychometric analysis. Under Rasch analysis, for students taking the same test form, the same raw score maps to the same logit measure. This means that all students who achieve 50% on the same test form are estimated as having the same ability, for example. This one-to-one correspondence of raw score to vocabulary size is a fundamental assumption of Nation and Beglar's (2007) definition of vocabulary size, a definition that requires each word to carry equal weight. This condition is satisfied by the Rasch model but not by more complex IRT models whose fundamental rationales are that items should not carry equal weighting (DeMars, 2010). This property of the Rasch model also simplifies linking of alternate test forms to a reference form through the use of score tables that criterion reference raw scores from each alternate form to vocabulary sizes estimates from the reference form.

Contextual effects and many-faceted Rasch measurement

Henning (1992) distinguished between *psychological* and *psychometric* unidimensionality. The former means that scores are interpretable in terms of the intended construct and the latter reflects homogeneity of item variances. Chapelle (1998) identified *trait, behaviorist,* and *interactionalist* models of knowledge. Trait models attribute knowledge to learner factors without specification of context. Nation (2012), for example, asserted that the VST tested vocabulary without context. Behaviorist views hold that knowledge can only be defined with reference to the context of use, while interactionalist models hold that both traits and contexts of use must be defined. Investigations of the effect of task type on item difficulty implicitly assume an interactionalist model of knowledge, where item difficulty derives from interaction of the word (the trait component) with the task type (the context), echoing Oller's argument that "knowing a word is knowing how to use it in a meaningful context" (1979, p. 189). One concern that arises here is that some VST item stems used a definitional sentence, requiring syntactic parsing, while others used a single word synonym. Nation recognized that "the difficult grammar of English definitions" (2012, p. 4) was problematic for the construct definition, so recommended the use of bilingual test forms. However, bilingual forms are problematic for any groups of students with varied L1s because they will not function as parallel test forms unless all the items, including the distractors, function identically in every test form. The mixing of multi-word and single word items in the VST thus raises questions about the psychological unidimensionality of the two item types, but Beglar (2010) reported sufficient psychometric unidimensionality that any sub-dimension related to syntactic parsing was not of major concern. Nation's (2012) advocacy of bilingual test forms is thus both unnecessary and undesirable if Beglar's (2010) analysis is accepted.

A further issue relating to unidimensionality concerns nuisance dimensions; small sub-dimensions that manifest differently in different contexts or at different times (Luecht & Ackerman, 2018). For example, Japanese language proficiency would constitute a nuisance dimension if foreign students in Japan were administered a bilingual vocabulary test that tested the synonymy of English and Japanese words. Test scores would represent a multidimensional trait of knowledge of both English and Japanese rather than a unidimensional trait of English knowledge. Nuisance dimensions are of particular concern for longitudinal studies, where pre-test and post-test scores may represent different composite constructs because of context related changes in nuisance dimensions. For example, foreign students studying at Japanese universities are often required to take both English and Japanese language classes so score gains on a bilingual vocabulary test administered at the beginning and end of a semester might represent improved Japanese proficiency as well as improved English proficiency.

A common approach to longitudinal datasets is to "rack" the data so that the pre-test and post-test responses for each item are analyzed as two separate items within a single analysis (Wright, 2003, p. 905). Tests of dimensionality and data-model fit can then be performed to investigate possible nuisance dimensions. However, this procedure violates the requirement of local item independence, so many-faceted Rasch measurement (MFRM) (Linacre, 1994) addresses this by allowing contextual variables to be modeled as measurement *facets* in addition to the familiar facets of *items* and *persons.* Rather than treating multiple responses by the same person to the same item as representing two items, MFMR treats them as one person responding to one item in different contexts. Although commonly used to model the effect of human raters in performance tests, such as in McNamara's (1996) seminal work, MFRM is applicable to any dataset where each student can interact with each item under different contextual conditions.

Background to this study

This study reanalyzed data collected at two Japanese universities, a public women's university and a private co-educational university. The introduction of a new Academic English Program (AEP) at the women's university led to disappointment when the expected TOEFL score improvements were not achieved, leading to curriculum reform and placement test development projects. One major issue was determining a suitable lexical level for both instructional and assessment content, consistent with Nation's (2012) recommended uses of the VST. It was also desirable to gather longitudinal data to determine whether the revised curriculum led to the intended improvements in language ability. A similar situation occurred at the co-educational university, where a proposed new language program raised questions about an appropriate level of content. The existing official course objectives assumed a level of proficiency that both Japanese and non-Japanese teachers considered unrealistic so criterion-referenced measures of the range of student ability were desirable to make recommendations for the proposed new program. To avoid detailed technical explanation about the use of logits and Rasch analysis, results were rescaled to vocabulary size estimates. The vocabulary sections of classroom and semester final tests used an item format based on the *Test of Vocabulary Synonymy* (TVS) used by Holster and Lake (2016), so these tests were linked and rescaled to the VST vocabulary size scale. As the test linking was conducted through Rasch analysis of concurrently administered items from the VST and TVS item banks, the essential research questions revolve around whether the requirements of the Rasch model were satisfied and claims of measurement invariance warranted.

Research questions

RQ1: Do rescaled logit scores and guessing-corrected raw scores provide invariant estimates of vocabulary size?

RQ2: Do the TVS and VST items measure a unidimensional construct?

RQ3: Are item difficulties from longitudinal datasets sufficiently invariant to support measurement of learning gains?

RQ4: Does item position within a test form affect measurement invariance?

Method

Participants

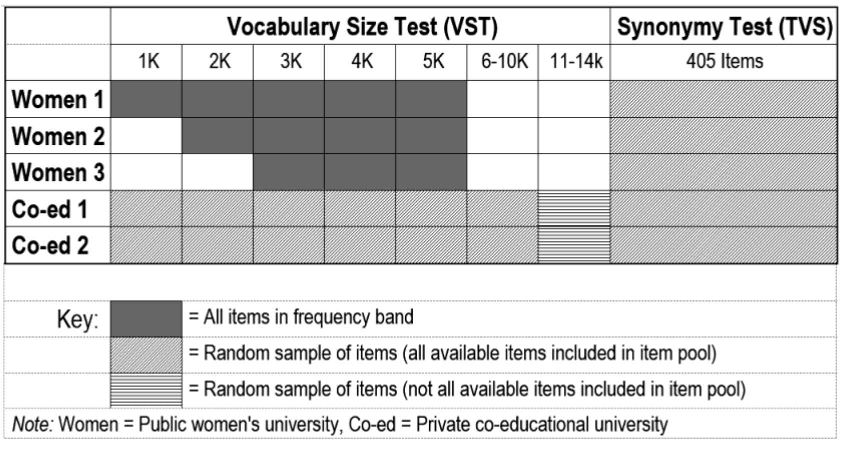
Tests were administered to a convenience sampling of 1,872 first-year students (typically 18 or 19 years old) taking compulsory English classes at a public women's university and a private co-educational university. Students came from a range of departments at each institution. Consistent with Beglar's (2010) sample of Japanese undergraduate students, students predominantly had vocabulary sizes below the 5K level.

Instruments

The 4-option VST provided a reference form for test linking and rescaling. A 50-item VST test was used at the women's university, limited to items in the 1K to 5K range, reflecting the typically low vocabulary sizes of Japanese students also noted by Beglar (2010). Some students were tested on all 50 items while others were administered 30-item or 40-item tests due to constraints on class time. Fifteen 50-item VST forms were created for use at the co-educational university, using items from the 1K to 14K range. Microsoft Excel was used to randomize item placement but biased to favor high-frequency words and to place them earlier in the test form. This algorithm resulted in inclusion of all items from the 1K to 10K bands, but gaps in the 11K to 14K range. The test administration pattern is shown schematically in Figure 1.

Figure 1

Test administration pattern



	Vocabulary Size Test (VST)							Synonymy Test (TVS)
	1K	2K	3K	4K	5K	6-10K	11-14k	405 Items
Women 1	■	■	■	■	■			▨
Women 2		■	■	■	■			▨
Women 3			■	■	■			▨
Co-ed 1	▨	▨	▨	▨	▨	▨	▤	▨
Co-ed 2	▨	▨	▨	▨	▨	▨	▤	▨

Key: ■ = All items in frequency band

▨ = Random sample of items (all available items included in item pool)

▤ = Random sample of items (not all available items included in item pool)

Note: Women = Public women's university, Co-ed = Private co-educational university

Note. Multiple test forms were created for both the VST and TVS, with quasi-random item placement. Students were administered 30 to 50 VST items and 108 TVS items as a pre-test and 108 TVS items as a post-test. Some students took a 54 item TVS test as a mandated final exam.

The 5-option TVS items used single-word synonyms rather than the definitional sentences of the VST to eliminate the syntactic parsing that Nation (2012) reported as problematic in the VST. TVS specifications (Holster & Lake, 2016) were used to develop additional items based on classroom materials by substituting synonyms for target words in listening and reading texts. For example, students heard the following sentences in one of the dialogues from a coursebook:

A: You're working? What do you do?

B: I'm a tutor.

However, in the transcript given to students, the target word *tutor* was changed to *teacher* and students were required to highlight any such discrepancies while they listened. Typically, 10 to 15 synonyms were presented each week. These were then tested in weekly written classroom review tests and semester tests, both contributing to a significant proportion of course grades. Based on the word frequencies published by Davies and Gardner (2010), the higher frequency synonym was used as the item stem and the lower frequency synonym as the key. Four distractors were selected from Davies and Gardner's (2010) list by finding the two next more frequent and two next less frequent words of the same part of speech as the stem and key, with any potentially problematic distractors skipped in favor of the next more or less frequent word. Students were instructed to read the item stem and identify a synonym from the five answer choices.

A sample test item is:

Teacher

A) Fee B) Tutor C) Sense D) Market E) Nation

Procedure

The VST was administered at the beginning of the first semester at the women's university to calibrate placement tests and at the end of the semester at the co-educational university to calibrate achievement tests. The original 50-item TVS formed the vocabulary section of the placement test used at the women's university. The entire placement test was administered again at the end of the semester, one week before the semester final test. At the co-educational university, new textbook derived TVS items were administered as weekly review tests, with the primary intention of rewarding students for being engaged in class and reviewing class materials each week. Weekly review test data was not included in this study. In order to familiarize students with the item formats used in the weekly review tests, a practice test was administered in the first or second week of class, with TVS items forming the vocabulary section. TVS items were administered again as the vocabulary section of the semester final test in the final week of class. Some courses were required by the co-educational institution to be administered an official final exam, typically two weeks after the end of the 15-week semester. These official final exams were administered by university staff under test conditions, were limited to one side of an A4 sheet of paper for administrative convenience, and could not include listening tasks because students taking different courses with different teachers were combined within each test room. These constraints limited the final exam to 54 TVS items, in contrast to the classroom administrations which contained 108 items on two A4 pages. Each TVS form was generated from an Excel workbook that randomized item placement. The 658 students in these courses thus took three administrations of the TVS. Test forms were scanned using Remark Office OMR version 8.4 and data analyzed with Winsteps version 4.0.0 and Facets version 3.8.04 using the default settings for the Rasch dichotomous model.

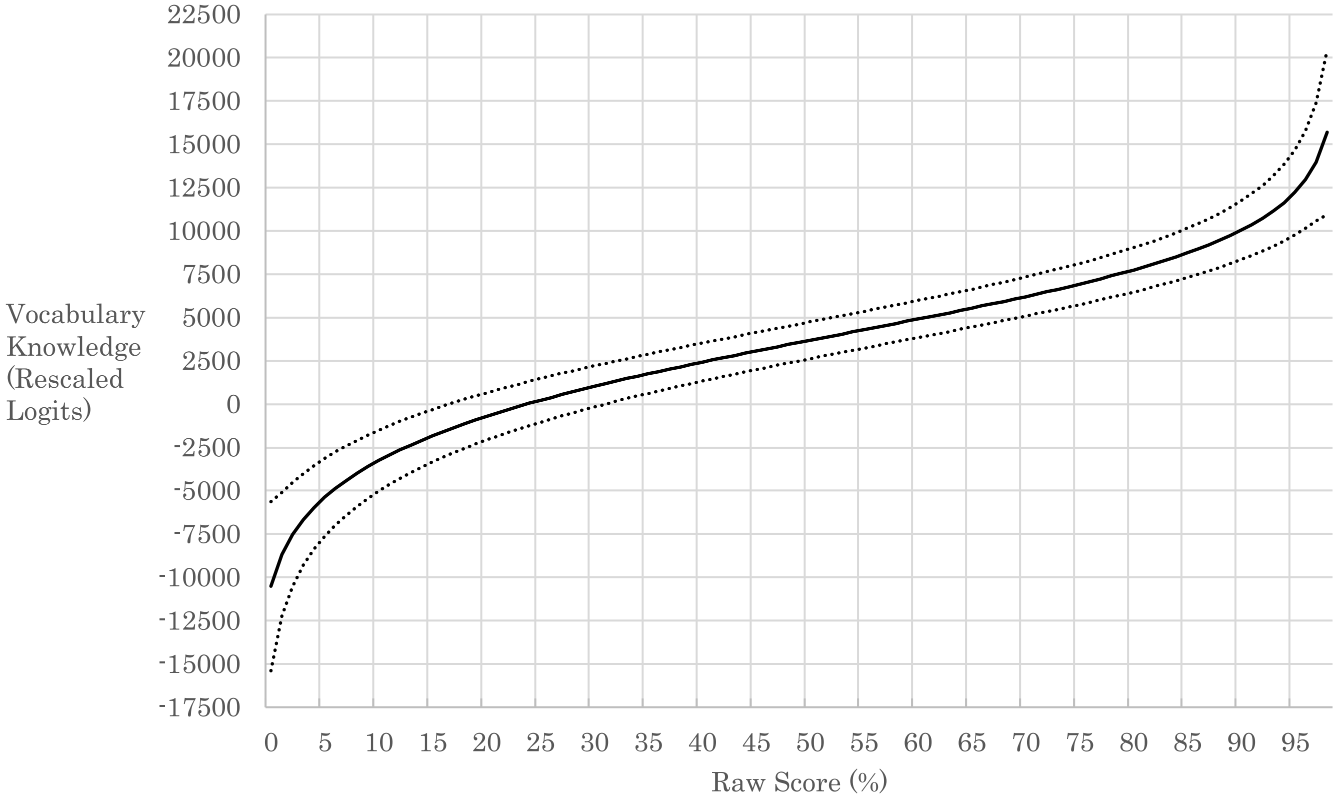
Results

Rescaling logit measures to vocabulary size

The first stage of analysis focused on rescaling item difficulty to a vocabulary size scale. Winsteps was used to produce a score table matching raw scores to logit measures for a VST reference form containing all 100 items in the 1K to 10K bands (hereafter VST10). Scaling of logit measures to vocabulary size was based on the following assumptions: 1) Mean item difficulty should be approximately 3,333 words, equaling the guessing-corrected vocabulary size of a person scoring 50%; 2) 1 logit should be scaled to 2,300 words, giving a 4 logit range from -1 logit (27%) to 3 logits (95%) corresponding to guessing-corrected vocabulary sizes of 67 words to 9,333 words. In practice, the relationship between logit measures and raw percentages was found to be approximately linear from raw scores of 25% to 80%, but increasingly non-linear beyond that. Empirical results showed that scaling 1 logit to 2,400 words, with mean difficulty of 3,300 words, produced a score table with close approximations between guessing-corrected raw scores and rescaled logits between 25% and 80%. These results are shown in Figure 2, with raw scores of 25% and 80% respectively producing VST sizes of approximately zero and 7,500 words, very close to the expected values. Rescaled logit scores are thus usefully invariant with vocabulary size estimates within the range of 0 to 7,500 words, allowing learning gains for students within this range to be expressed in terms of word families known.

Figure 2

Raw VST10 score versus vocabulary knowledge



Note. The upper and lower dashed lines show the 95% confidence intervals. The vertical axis shows logit scores rescaled to estimated vocabulary size, with 1 logit = 2 400 words.

Figure 2 also shows 95% confidence intervals, typically spanning a range of about 2,500 words, evidence that the VST10 is unsuitable for measurement of individual student learning unless very large learning gains have been achieved. This is not a reflection on the VST's validity as a general measure of vocabulary sizes, it simply reflects that it was not intended to be precise enough to measure small learning gains by individual students. Figure 3, mapping person ability against item difficulty, confirms this, with mean person ability of 2,113 words and a standard deviation of 1,654 words. The confidence interval is thus about 1.5 standard deviations of this sample of persons, meaning that more items are required to reduce the measurement error. Figure 3 also shows many items that were far too difficult for any student, so measurement quality would be improved by removing items above the 5K level and replacing them with 1K and 2K items. Additionally, although the VST sampled equally across frequency bands, the distribution of item difficulties did not reflect this, confirming the presence of many idiosyncratic items observed by Beglar (2010). An important implication of this finding is that the suitability of items for many classroom testing purposes will be determined by the empirically derived logit difficulty rather than the BNC frequency band, whereas researchers may prefer to select items based on frequency to simplify estimation of vocabulary size. Comparison with Figure 2 shows that the highest density of item difficulty aligns with the range of vocabulary sizes that show the most linear relationship with logit measures. Clearly, many more items were required in the 1K and 2K bands and many fewer items above the 5K level were needed, a limitation the TVS items were developed to address.

Figure 3

Person-item map of VST10 results

```
MEASURE                                 |                               MEASURE
  <more> --------------------- Persons -+- items   ----------------- <rare>
 9600                                   +T X                             9600
 9000                                   +                                9000
 7800                                   +  X                             7800
 7200                                .  +  XXXXXXX                       7200
 6000                                .  +S XXXXXXXXXXX                   6000
 5400                              .## T+  XXXXXXXXXXXX                  5400
 4800                          .######  +  XXXXXXXXXXXXXXX               4800
 3600                .################ S+M XXXXXXXXXXXXXX                3600
 3000       .#########################  +  XXXXXXXX                      3000
 1800      .########################## M+  XXXXXXXXX                     1800
 1200                 .###############  +  XXXXXX                        1200
  600                     .########### S+S XXX                            600
 -600                           ######  +  X                             -600
-1200                               .# T+  XXXX                         -1200
-2400                               .#  +  X                            -2400
-3000                                .  +T XX                           -3000
-3600                                .  +  XX                           -3600
-4800                                   +                               -4800
-5400                                   +  XX                           -5400
-6600                                .  +  X                            -6600
-7200                                   +                               -7200
  <less> --------------------- Persons -+- items   ----------------- <freq>
 Each "#" in the Persons column is 5 Persons: Each "." is 1 to 4
```

Note. Persons and items are mapped against a common scale of vocabulary knowledge expressed as words known. Higher placement on the map indicates higher person ability or higher item difficulty.

Linking VST and TVS test forms

RQ2 addressed the unidimensionality of the VST and TVS items, a fundamental requirement for linking the two tests. Table 1 shows principal components analysis of residuals (PCAR) from the combined VST and TVS dataset. The Rasch dimension explained 35.6% of total variance, exceeding Reckase's (1979) guideline of a minimum of 20% variance explained, with the largest subdimension accounting for 0.6% of variance. These results should be treated cautiously due to the low data density but are consistent with the TVS and VST measuring a unidimensional construct.

Table 1

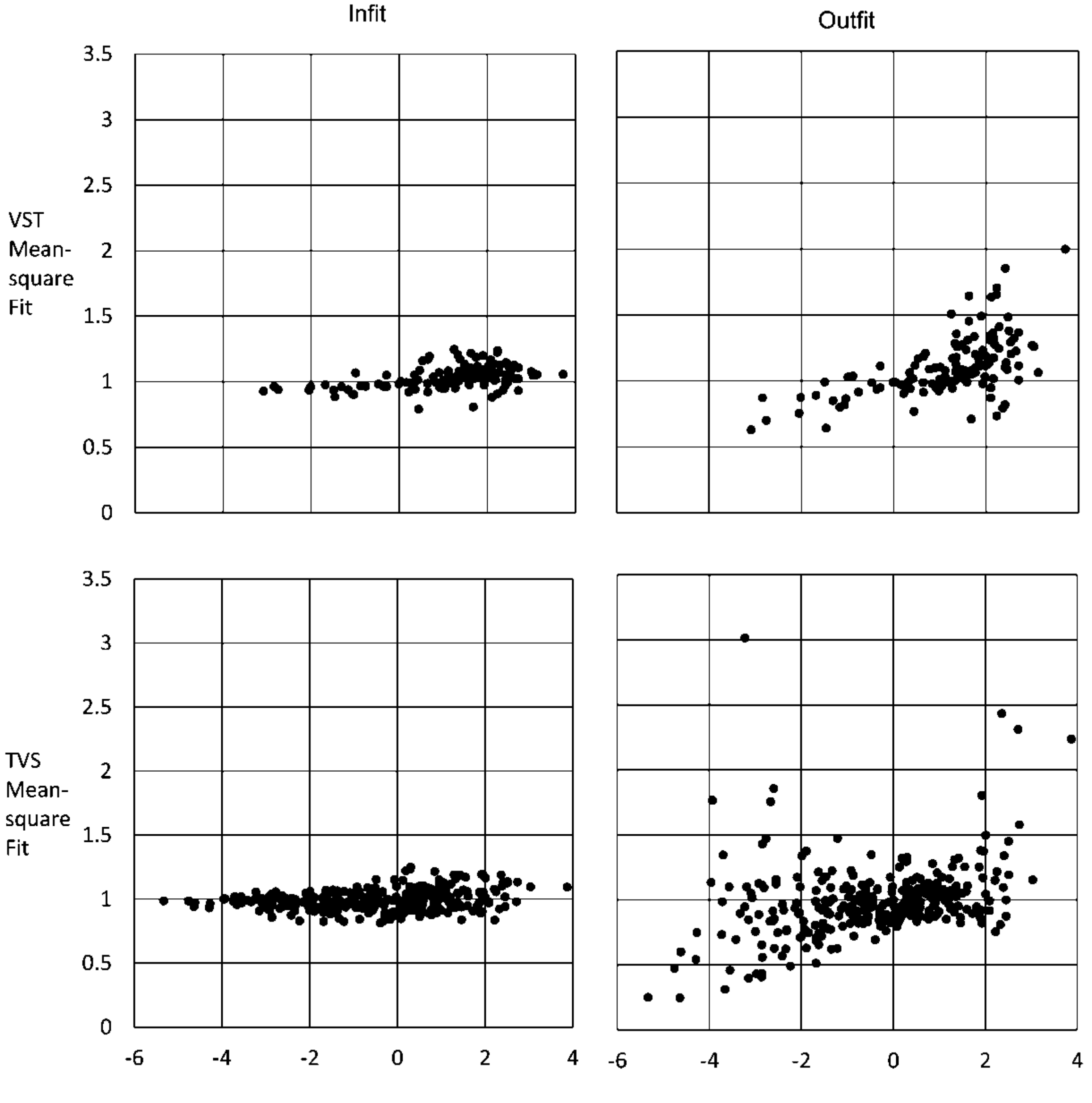
Variance explained by measures

Variance		Eigenvalue	Observed %	Expected%
Total:		655.7	100.0%	100.0%
Rasch:	Measures	233.7	35.6%	35.6%
	Persons	61.7	9.4%	9.4%
	Items	172.0	26.2%	26.2%
Unexplained:	Total	422.0	64.4%	64.4%
	1st contrast	4.2	0.6%	
	2nd contrast	3.8	0.6%	
	3rd contrast	3.5	0.5%	

Dimensionality can also be investigated by checking for systematic patterns in mean-square fit statistics for the VST items and TVS items, as shown in Figure 4. The VST items, shown in the two upper panels, were more difficult on average than the TVS items, shown in the two lower panels, with very few VST items below 0.00 logits compared with many TVS items. Mean-square infit, reflecting information weighted responses, is shown in the two left-hand panels, with all values comfortably below Linacre's (2009) 1.50 rule-of-thumb guideline for concern. Mean-square outfit, reflecting unweighted response, is shown in the two right-hand panels, with difficult items tending to misfit for both item types. Although easy items of both types showed a tendency to overfit, this pattern was very pronounced for the VST items. The easy TVS items were somewhat less consistent, with an extreme range of outfit including some highly overfitting items and some highly misfitting items, but there were insufficient easy VST items to draw firm conclusions. The information-weighted infit mean-square value is a crucial indicator of measurement quality (Linacre, 2009), and all items performed acceptably. The outfit mean-square value indicates unexpected outlying responses, with 17 items exceeding the 1.50 threshold of concern, including four TVS items with values exceeding 2.0. In a battery of 422 items, 17 misfitting items constitutes about 4% of total items and all the misfitting items were at the extremes of the measurement range so these do not pose a substantive threat to test linking.

Figure 4

Mean-square item fit



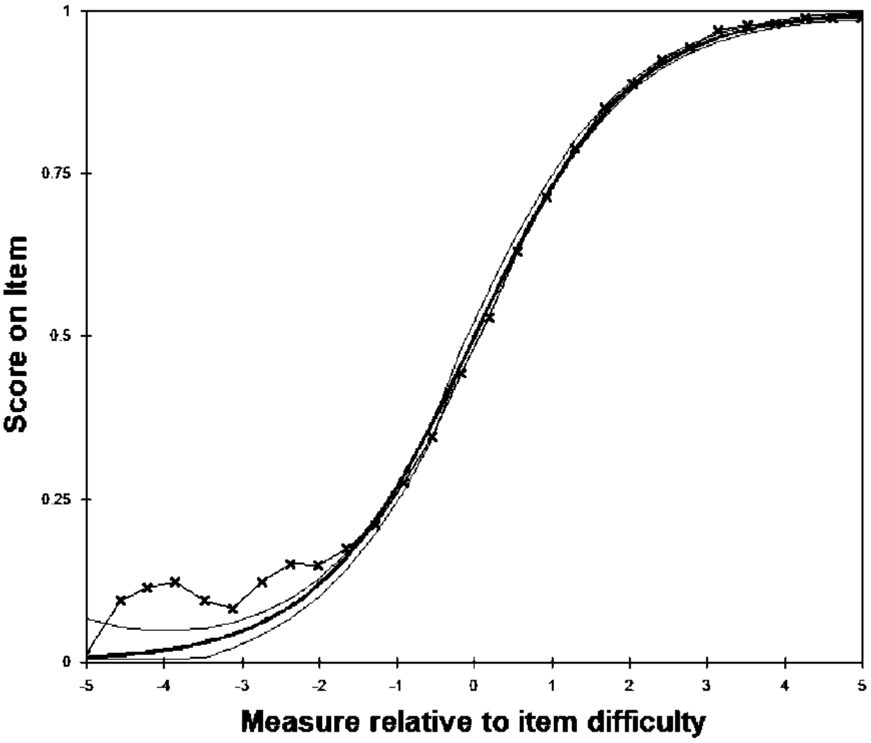
Item Difficulty (Logits)

Note. The two upper panels show VST items, the two lower panels show TVS items. The horizontal scale shows item difficulty in logits, the vertical scale shows mean-square fit, with infit shown in the two left-hand panels and outfit in the two right-hand panels. Mean-square values lower than 1.50 indicate acceptable item functioning.

Figure 5 shows the modelled and empirical test characteristic curves for the combined VST and TVS analysis. The empirical results closely match the Rasch model above a probability of success of approximately 20%, below which the results misfitted the model. These results are consistent with low-ability persons succeeding on difficult items through random guessing, with odds of random guessing of 25% on the VST items and 20% on the TVS items. This illustrates the importance of S.R. test items being well matched to the ability of test takers. Figure 5 supports the view that the misfit associated with difficult items in Figure 4 arose due to random guessing but does not resolve the cause of the misfit of easy items. Item dependency was analyzed through correlations between standardized item residuals, the standard Rasch procedure (Aryadoust et al., 2021), with the 10 largest values shown in Table 2. One pair of items showed a correlation of .76, meaning that shared variance exceeded 50%, the level at which inter-item dependency exceeds random variance (Linacre, 2020). These two items tested the synonymy of *good/nice* and *child/youngster,* the two items having no obvious semantic relationship. One other pair of items showed a correlation of .66, indicating 44% shared variance. These items tested *talk/speak* and *seafood/fish,* which also have no obvious semantic connection. Given the lack of semantic connection, these four items do not threaten the requirement of independence in a test battery of over 400 items.

Figure 5

Empirical versus modelled test characteristic curve for combined VST and TVS items



Note. The solid central line shows the modelled expectation of success for persons of different ability, with each X showing observed probabilities and the upper and lower solid lines showing confidence intervals.

Table 2

Standardized residual correlations

Item Number	*Item Number*	*Correlation*
196	400	.76
151	411	.66
130	368	.52
122	126	.52
130	403	.51
343	365	.45
126	130	.41
71	89	.40
256	262	.40
183	267	.40
368	403	.39
106	119	.39
115	120	.37

Table 3 shows fit statistics for the 17 items with mean-square values exceeding 1.50. All had low or negative point-measure correlations, indicating an inability to discriminate between high and low-proficiency persons. Twelve of the items were extremely difficult, with logit values exceeding 1.90, corresponding to vocabulary sizes exceeding 7,500, and raw scores within the range of random guessing. Three misfitting items were extremely easy, with five incorrect responses or fewer, meaning that a single careless response would be sufficient to cause misfit. The remaining two misfitting items were the VST items *Gimmick* and *Upbeat*, with respectively 11/38 and 18/50 correct responses and outfit mean-square values of 1.64 and 1.51. Table 3 provides further evidence that misfit resulted from a very small number of responses so a larger sample of persons would likely have resulted in better fit (and point-measure correlations). This small number of misfitting responses does not pose a substantive threat to test linking because the large number of well-fitting responses included all the items matched to the range of person ability. These well-matched items provide much more information than the outlying items, reflected in the much lower levels of infit than outfit. In response to RQ2, PCAR analysis and data-model fit indicated sufficient unidimensionality to map VST and TVS items into a common measurement scale for the purpose of measuring score gains across a semester of instruction.

Table 3

Most misfitting items

Item	Freq.	Synonyms					Infit		Outfit		Pt-M
Number	Level	Tested	Count	Score	Logits	*SE*	*MS*	*ZStd*	*MS*	*ZStd*	Corr
323	S3:	Help-Assist	115	112	-3.23	0.59	1.05	0.28	3.01	2.03	-.08
417	S5:	Potential-Implied	118	24	2.36	0.24	1.19	1.39	2.43	5.04	-.10
343	S3:	Gradually-Slowly	114	16	2.71	0.28	0.98	-0.05	2.31	3.58	.11
357	S4:	Compose-Write	231	12	3.87	0.30	1.09	0.44	2.23	2.67	-.11
49	V5:	Fracture-Break	207	12	3.72	0.30	1.06	0.31	2.00	2.46	-.08
80	V8:	Mumble-Speak	88	15	2.42	0.29	1.15	0.83	1.85	2.93	-.16
183	S1:	Girl-Daughter	112	107	-2.60	0.46	1.02	0.17	1.85	1.40	.00
365	S4:	Genuine-Actual	114	29	1.92	0.22	1.19	1.74	1.80	3.95	-.04
144	S1:	School-University	335	331	-3.93	0.50	1.00	0.17	1.76	1.22	.03
148	S1:	Look-Watch	114	109	-2.67	0.47	1.00	0.13	1.75	1.29	.06
117	V12:	Coven-Society	13	2	2.23	0.79	1.23	0.58	1.70	1.09	-.22
108	V11:	Hutch-Cage	52	10	2.23	0.36	1.24	1.12	1.66	2.12	-.30
68	V7:	Gimmick-Trick	38	11	1.63	0.37	1.08	0.57	1.64	2.47	-.01
77	V8:	Locust-Insect	75	15	2.12	0.30	1.13	0.79	1.63	2.45	-.11
319	S3:	Column-Tower	230	34	2.74	0.19	1.13	1.06	1.58	2.66	.02
95	V10:	Upbeat-Good	50	18	1.25	0.31	1.25	2.20	1.51	2.94	-.12
305	S3:	Purchase-Invest in	113	25	2.01	0.24	1.16	1.27	1.50	2.49	-.02

Note: VST items are coded "V" followed by frequency band. TVS items are coded "S" followed by frequency band. Count = number of responses recorded; Score = number of correct responses; Pt-M Corr = Point-measure correlation.

Linking longitudinal data using MFRM

Measuring learning gains through pre-tests and post-tests introduces a potential problem of multi-dimensionality due to nuisance dimensions, which may not be detected by tests of unidimensionality commonly used in IRT analysis (DeMars, 2010). MFRM allows time of administration to be isolated as a separate measurement facet and fit statistics to be analyzed for evidence of measurement distortion due to contextual effects. Longitudinal data was analyzed using a 4-faceted model using Facets version 3.80.0, with the facets of *Time* and *Position* added to the usual facets of *Persons* and *Items*. *Time* refers to the time of administration of the test; the beginning of the course (Week 1), the final class (Week 15), or during the official exam period (Final Exam). *Position* refers to the location of the item in the test form, ranging from 1 (the first item) to 108 (the final item). Responses from all 1,872 persons were used to calibrate the TVS items to the VST10 scale, including 24 VST items from the 11K to 14K bands, giving 529 items in total. This calibration was achieved through concurrent equating, with mean item difficulty adjusted empirically so the average difficulty of the VST items remained constant. Item difficulties from the combined analysis were then compared with those from the VST and TVS datasets analyzed in isolation. Summary statistics are shown in Table 4, with the mean of all items found to be -366 words and the TVS items to be -1737 words. This range represented a difference between the mean VST10 and TVS item of 2.10 logits, or 5,036 words on the VST10 scale.

Table 4

Summary statistics of VST and TVS items

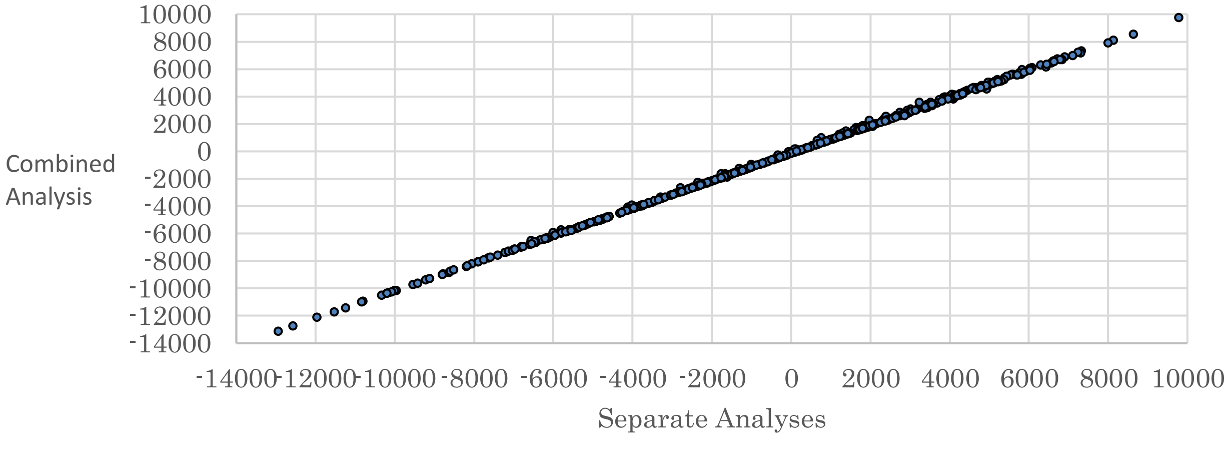
Item			Difficulty		Mean	Fair		Infit	Outfit	Pt-M	Item
Subset		*n*	(Words)	*SE*	Score	Ave	Count	*MS*	*MS*	Corr	Rel
All:	*M*	529	-366	598	0.64	0.61	549.2	1.01	1.02	.25	.96
	S.D.		4436	650	0.27	0.30	562.4	0.09	0.29	.16	
VST:	*M*	124	3299	519	0.41	0.39	236.4	1.01	1.03	.21	.97
	S.D.		3156	248	0.23	0.24	183.7	0.07	0.17	.15	
TVS:	*M*	405	-1737	571	0.72	0.70	657.3	0.99	0.98	.28	.95
	S.D.		4055	698	0.24	0.25	594.1	0.09	0.29	.14	

Note: Count = number of responses recorded; *Mean Score* = proportion of correct responses: *Fair Ave* = Probable mean score if all persons attempted all items; *Pt-M Corr* = Point-measure correlation; *Item Rel* = Reliability of item separation.

Figure 6 compares item difficulties for the combined and separate analyses of VST and TVS items, with deviations from the linear trendline much smaller than the typical measurement errors of 500 words shown in Table 4. Item reliability of all three analyses exceeded .95, indicating a stable hierarchy of item difficulty. The mean score column in Table 4 shows the observed average score, while the fair-average column shows the expected score if all students had taken all items. Clearly, the TVS items were much easier than the VST items, with respective fair-average scores of 70% versus 39%. This is consistent with the TVS items being targeted at the 5K frequency band and lower. Noteworthy is that the TVS items were slightly overfitting on average and had a higher point-measure correlation at .28 compared with .21 for the VST items, reflecting the better match of item difficulty to student ability. These results provide evidence that the combination of cross-sectional VST data and longitudinal TVS data was not a threat to measurement invariance, addressing RQ3.

Figure 6

Item difficulty for combined analysis versus separate analyses



Note. All 529 items were analyzed together for the combined analysis. For the separate analyses, the mean item difficulty of each sub-set of items was set to the value obtained from the combined analysis.

Measuring learning gains

The VST10 anchored item difficulties were then used to analyze learning gains for the 658 students at the co-educational university who were administered the TVS as an official final exam. This calibration allowed comparison between the classroom test in the final week with the final exam one or two weeks later. TVS items with respective infit and outfit mean-square fit values below 1.20 and 1.30 were anchored to the VST10 scale, with less-fitting items unanchored to avoid measurement distortion. Learning gains were then measured against this anchored scale, shown in Figure 7, with the estimated VST10 vocabulary size shown on the left. Students showed a substantive gain between Week 1 and Week 15, and also between Week 15 and the final exam. Summary statistics for all four facets are shown in Table 5. Although average mean-square statistics were close to the expected value of 1.00, *Persons* and *Items* had outfit standard deviations of 0.35 and 0.32 respectively, consistent with the misfit to outlying responses discussed earlier.

Table 5

Summary statistics of measurement facets

Facet		*N*	Rel	Sep	Count	Mean Score	Fair Ave	Vocab Size	*SE*	Infit *MS*	Outfit *MS*
Persons:	*M*	658	.95	4.48	314.60	.72	.79	2088	395	1.01	1.01
	SD				94.70	.11	.12	1860	92	0.11	0.35
Time:	*M*	3	1.00	28.56	69013.00	.74	.81	0	28	1.01	0.98
	SD				30320.10	.07	.05	823	10	0.02	0.08
Position:	*M*	108	.96	5.20	1917.00	.72	.81	0	154	1.01	0.99
	SD				76.50	.11	.06	824	24	0.05	0.14
Items:	*M*	380	.95	4.16	544.80	.71	.73	-1518	594	0.98	0.97
	SD				484.20	.24	.23	3968	714	0.17	0.32

Note: Count = number of responses recorded; *Mean Score* = proportion of correct responses: *Fair Ave* = Probable mean score if all persons attempted all items.

Figure 7

TVS facets measurement rulers

```
Measr |+Persons    |+Time       |-Position  |-items
   +------+------------+------------+-----------+------------+
    10200 +            +            +           +
     9600 +            +            +           +
     9000 +            +            +           +
     8400 +            +            +           + .
     7800 + .          +            +           + .
     7200 + .          +            +           + .
     6600 + .          +            +           + .
     6000 + .          +            +           + .
     5400 + *.         +            +           + *
     4800 + ***.       +            + .         + ****
     4200 + ****.      +            +           + ***.
     3600 + ******.    +            +           + ***.
     3000 + *******.   +            +           + ****.
     2400 + *********. +            + .         + *******
     1800 + ********.  +            + .         + *******
     1200 + ******.    +            + *         + *****.
      600 + *****.     + Final Exam + **.       + ******.
    *   0 * ***.       * Week 15    * ********* * ******.
     -600 + ***.       + Week 1     + ***       + *****.
    -1200 + **.        +            + *         + ********.
    -1800 + *          +            + .         + **********
    -2400 + .          +            +           + ********
    -3000 + .          +            +           + *****
    -3600 +            +            +           + ****.
    -4200 +            +            +           + ******.
    -4800 + .          +            +           + ***.
    -5400 +            +            +           + ***.
    -6000 +            +            +           + *****.
    -6600 +            +            +           + ****
    -7200 +            +            +           + ***.
    -7800 +            +            +           + *.
    -8400 +            +            +           + *.
    -9000 +            +            +           +
    -9600 +            +            +           + *
   -10200 +            +            +           + ****.
   +------+------------+------------+-----------+------------+
    Measr | * = 10     |+Time       | * = 6     | * = 3
```

Note. The measurement scale on the left is calibrated to VST10 vocabulary sizes. The Time facet shows gains through a 15-week semester. The position facet shows a substantively large effect of item position within the test forms.

Table 6 provides the measurement report for the *Time* facet, with a gain of 0.24 logits (570 on the VST10 scale) between Week 1 and Week 15, and a further gain of 0.44 logits (1,050 on the VST10 scale) between Week 15 and the final exam.

Table 6

Time measurement report

Time	Count	Mean Score	Fair Ave	Vocab Size	*SE*	Infit *MS*	*ZStd*	Outfit *MS*	*ZStd*	Disc	Pt-M Corr
Final	34452	0.82	0.87	891.8	39.0	1.01	1.3	0.91	-2.1	1.00	.44
Week 15	81450	0.72	0.81	-161.0	22.9	0.98	-3.6	0.95	-2.6	1.03	.53
Week 1	91137	0.69	0.77	-730.7	21.1	1.03	6.7	1.06	4.0	0.95	.53
M	69013.0	0.74	0.81	0.0	27.7	1.01	1.5	0.98	-0.3		.50
SD	30320.1	0.07	0.05	823.2	9.9	0.02	5.2	0.08	3.7		.05

Pop: RMSE 28.80 S.D. 671.5 Separation 23.31 Strata 31.42 Reliability 1.00
Samp: RMSE 28.80 S.D. 822.7 Separation 28.56 Strata 38.42 Reliability 1.00
Fixed (all same) chi-square: 1387.3 d.f.: 2 significance (probability): .00
Random (normal) chi-square: 2.0 d.f.: 1 significance (probability): .16

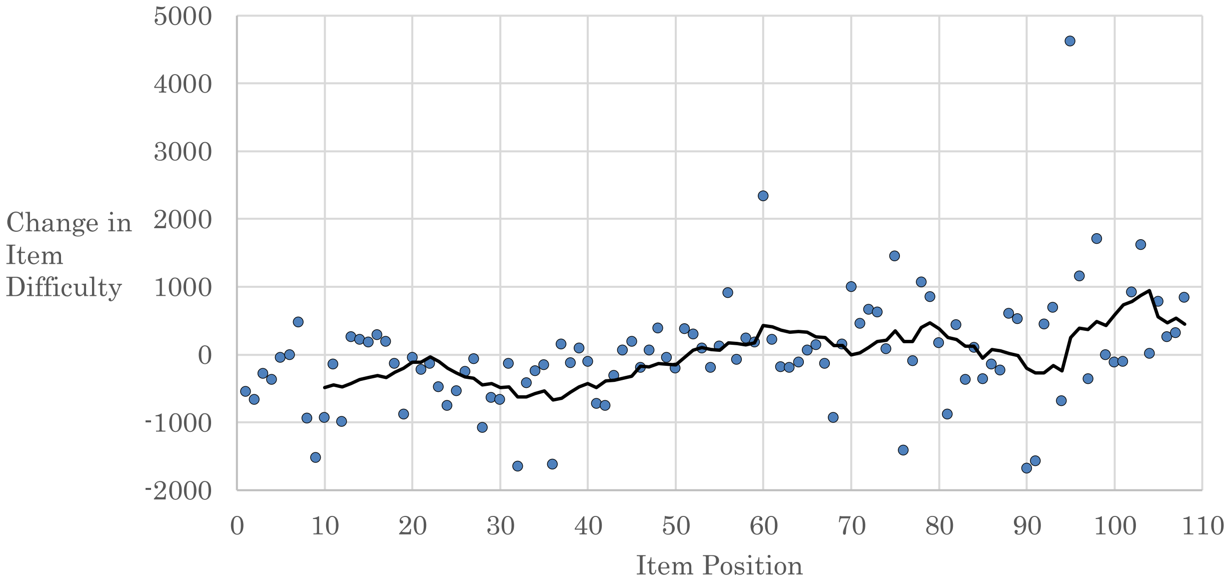
Note: Count = number of responses recorded; *Mean Score* = proportion of correct responses: *Fair Ave* = Probable mean score if all persons attempted all items; *Disc* = Discrimination; *Pt-M Corr* = Point-measure correlation.

Invariance of item position

Also of note from Figure 7 is that the *Position* facet has a substantively large range, with a standard deviation of 824 words (0.34 logits) reported in Table 5. Figure 8 shows item difficulty versus position, with each point on the solid trendline showing the mean of the preceding 10 items, allowing the general trend to be visible though the fluctuations in the data. Although the trend is quite noisy, moving an item from the beginning of a test form to the end would typically result in item difficulty increasing by the equivalent of 1,000 words. In a test such as the VST, with items ordered by frequency band, the difficulty of high-frequency items would be substantively under-estimated and the difficulty of low-frequency items overestimated, so research into the relationship between frequency and difficulty should take this effect into consideration.

Figure 8

The effect of item position on difficulty

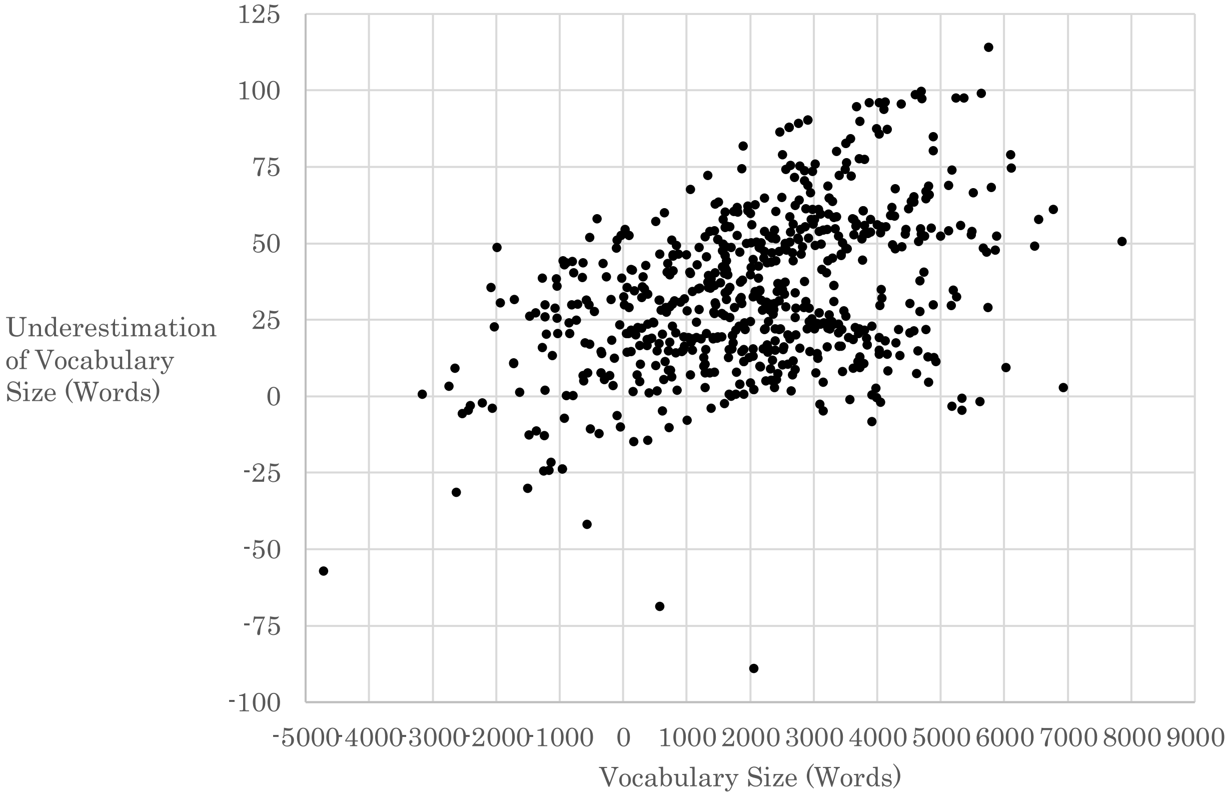


Note. The solid trendline shows the moving average of 10 items, with placement near the end of the test associated with a substantive increase in item difficulty.

In this study, however, the objective was the measurement of person ability rather than item difficulty and the use of a randomization algorithm greatly weakened the relationship between frequency and item position. Figure 9 shows the effect on vocabulary size estimates of including item position as a measurement facet, with the vertical axis scale exaggerated for emphasis. Person ability increased by an average of about 34 words when item position was included, with a greater effect on higher-ability persons. However, the substantive size of the effect is very small compared with the SE of 395 words reported in Table 5. The effect of item placement on the estimated vocabulary size of an individual student is thus an order of magnitude smaller than the measurement error, so not of concern to classroom teachers. A qualified answer to RQ4 is thus that item position within a test form has an effect too small to substantively affect the measurement of individual persons, but large enough to be of concern to researchers investigating the relationship between word frequency and item difficulty. It is therefore recommended that researchers include item position as a measurement variable.

Figure 9

The effect of including item position on estimates of vocabulary size



Note. The vertical axis shows the difference in vocabulary size after including item position as a measurement facet. Note that the scale of the two axes differs by two orders of magnitude.

Discussion

This study investigated the rescaling of classroom vocabulary tests to the VST scale using Rasch modelling. Although the VST was developed to provide a general indication of students' vocabulary sizes (Nation, 2012), it provides only 10 items per 1K frequency band. As the majority of students in this study had vocabulary sizes below the 3K level, relatively few VST items were matched to students' levels, limiting measurement precision and making it unsuitable as an instrument to measure learning gains. Synonymy test items based on textbook content were therefore developed to provide a much larger pool of items below the 5K level. However, the synonymy test was not designed to sample equally across all relevant frequency levels, a necessity for the estimation of vocabulary size using the protocol established by Nation and Beglar (2007). The VST was therefore administered as a reference test in order to calibrate the classroom tests to the VST scale using Rasch analysis.

The estimation of vocabulary size is based on raw scores which do not provide invariant interval level measurement, a major limitation on the potential use of scores. RQ1, the major research question, investigated invariance between guessing-corrected estimates of vocabulary size and logit scores, finding sufficient invariance for the purpose of test linking. This linking demonstrates how a measure of vocabulary size can be rescaled to a VST derived scale using vocabulary tests developed to different specifications.

RQ2 investigated the requirement that a unidimensional construct underlies both the VST and TVS despite the very different interpretations of the resulting scores. Unidimensionality is a requirement for the analysis of raw percentage scores as well as Rasch analysis. Although both the VST and TVS required students to match synonymous expressions, the VST included definitional phrases whereas the TVS used only single-word synonyms. The VST and TVS items were found to be consistent with a strongly unidimensional trait of vocabulary knowledge, supporting the appropriacy of test linking.

RQ3 investigated whether invariance was maintained across longitudinal data, a requirement for the measurement of learning gains. Item difficulty was found to be usefully invariant, evidence that any nuisance dimensions related to time of test administration were too small to effect test linking.

RQ4 investigated the effect of item position on difficulty. This study found a statistically significant effect whereby placement near the end of the test increased item difficulty. Although too small to be of concern for testing person ability, this effect threatens the validity of research into the relationship between word frequency and item difficulty. This is because it is standard practice to arrange test items in order of decreasing frequency, such as in the VST forms published by Nation and Beglar (2007) and Nation (2012). Although Beglar (2010) found a general tendency for lower frequency items to be more difficult, this effect was much more pronounced for very high-frequency items, with a very small effect above the 10K level. The effect of item position on difficulty observed in this study makes it plausible that Beglar's results exaggerated the effect of frequency on difficulty and that the very small increases above the 10K level actually reflected item position, not item difficulty itself. Although this is speculative given the different sampling of students and test administration protocols, it is a plausible hypothesis given the results found in this study. Researchers investigating the relationship between item difficulty and frequency need to either empirically demonstrate that item position does not affect item difficulty or use multiple test forms with randomized item placement.

Conclusion

This study demonstrated the use of Rasch modelling of vocabulary size, vocabulary size being an ordinal scale of measurement based on the protocol of assigning the same vocabulary size to students with the same raw score on the same test form. Conceptualizing vocabulary size as invariant carries the implication that scores from different test forms can be linked and calibrated to a common scale. The Rasch model provides this invariance and also supports the one-to-one mapping of raw score to vocabulary size that underlies the concept of vocabulary size. However, measurement invariance requires psychometric unidimensionality and acceptable data-model fit. This study found sufficient unidimensionality to rescale scores from a test of vocabulary synonymy and to measure gains over a semester. Raw scores are fundamentally unable to provide invariant vocabulary size estimates because of practical limits on test length. Decreasing test length by removing low-frequency items will cause underestimation of vocabulary size, a problem addressed through Rasch linking of test forms. Contextual effects are a further threat to the invariance of vocabulary size estimates, with item position shown to cause substantive misestimation of item difficulty.

References

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101-118. https://doi.org/10.1177/0265532209340194

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, *30*(4), 277-291. https://doi.org/10.1111/j.1745-3984.1993.tb00427.x

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, *10*(2), 157-187. https://doi.org/10.1177/026765839401000203

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). Cambridge University Press.

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English*. Routledge.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Dollerup, C., Glahn, E., & Hansen, C. R. (1989). Vocabularies in the reading process. *International Association of Applied Linguistics Review*, *6*, 21-33.

Engelhard, G. (2013). *Invariant measurement*. Routledge.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, *7*(2), 33-38. https://doi.org/10.1111/j.1745-3992.1988.tb00434.x

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1-11. https://doi.org/10.1177/026553229200900102

Holster, T. A., & Lake, J. W. (2016). Guessing and the Rasch model. *Language Assessment Quarterly*, *13*(2), 124-141. https://doi.org/10.1080/15434303.2016.1160096

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). MESA Press.

Linacre, J. M. (2009). *Misfit diagnosis: infit outfit mean-square standardized*. Retrieved 18 January from http://www.winsteps.com/winman/index.htm?globalfitstatistics.htm

Linacre, J. M. (2020). *Table 23.99 Largest residual correlations for items*. Retrieved 12 July from https://www.winsteps.com/winman/table23_99.htm

Luecht, R., & Ackerman, T. A. (2018). A technical note on IRT simulation studies: Dealing with truth, estimates, bbserved data, and residuals. *Educational Measurement: Issues and Practice*, *37*(3), 65-76. https://doi.org/doi:10.1111/emip.12185

McNamara, T. F. (1996). *Measuring second language performance*. Pearson Education.

Nation, I. S. P. (2012). *The vocabulary size test*. Retrieved 22 August from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9-13. http://jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf

Oller, J. W. (1979). *Language tests at school: A pragmatic approach*. Longman.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Denmark Paedogiske Institut.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, *19*(2), 12-25. https://doi.org/10.1177/003368828801900202

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511732942

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230. https://doi.org/10.2307/1164671

Santellices, M. V., & Wilson, M. (2010). Unfair treatment?: The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review*, *80*(1), 106-133. https://doi.org/10.17763/haer.80.1.j94675w001329270

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109-120. https://doi.org/10.1017/S0261444819000326

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680. https://doi.org/10.1126/science.103.2684.677

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, *26*(2), 161-176. https://doi.org/10.1111/j.1745-3984.1989.tb00326.x

Wright, B. D. (2003). Rack and stack: Time 1 vs. time 2 or pre-test vs. post-test. *Rasch Measurement Transactions*, *17*(1), 905-906. https://www.rasch.org/rmt/rmt171a.htm

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Appendix

Table A1

Randomization algorithms for VST test forms

Frequency	Randomization Algorithm
1K	=RANDBETWEEN(1,20000)
2K	=RANDBETWEEN(1000,20000)
3K	=RANDBETWEEN(2000,20000)
4K	=RANDBETWEEN(3000,20000)
5K	=RANDBETWEEN(4000,20000)
6K	=RANDBETWEEN(5000,20000)
7K	=RANDBETWEEN(6000,20000)
8K	=RANDBETWEEN(7000,20000)
9K	=RANDBETWEEN(8000,20000)
10K	=RANDBETWEEN(9000,20000)
11K	=RANDBETWEEN(10000,20000)
12K	=RANDBETWEEN(11000,20000)
13K	=RANDBETWEEN(12000,20000)
14K	=RANDBETWEEN(13000,20000)

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)

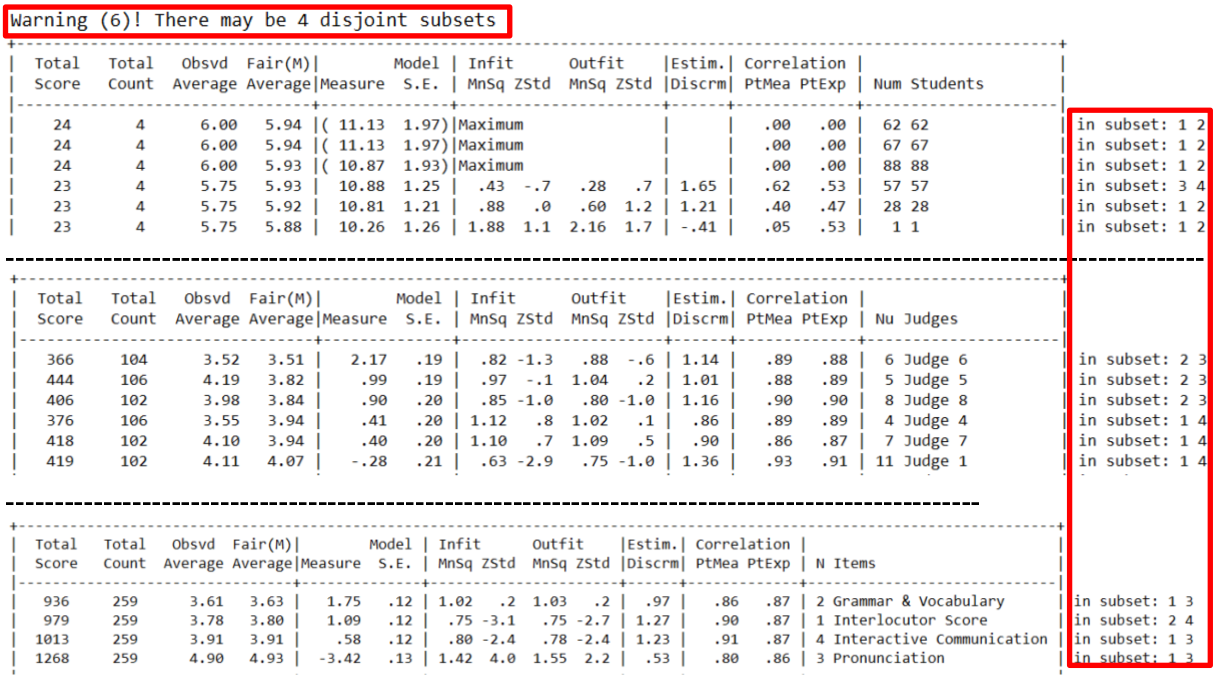
				Interlocutor	Rater		
Day	Student	Teacher	Categories	Holistic Score	Grammar + Vocabulary Score	Pronunciation Score	Interactive Communication Score
1	1	1	1-4a	5	#	#	#
	1	6	1-4a	#	4.5	4.5	5
2	15	8	1-4a	4.5	#	#	#
	15	1	1-4a	#	5	4.5	4.5

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

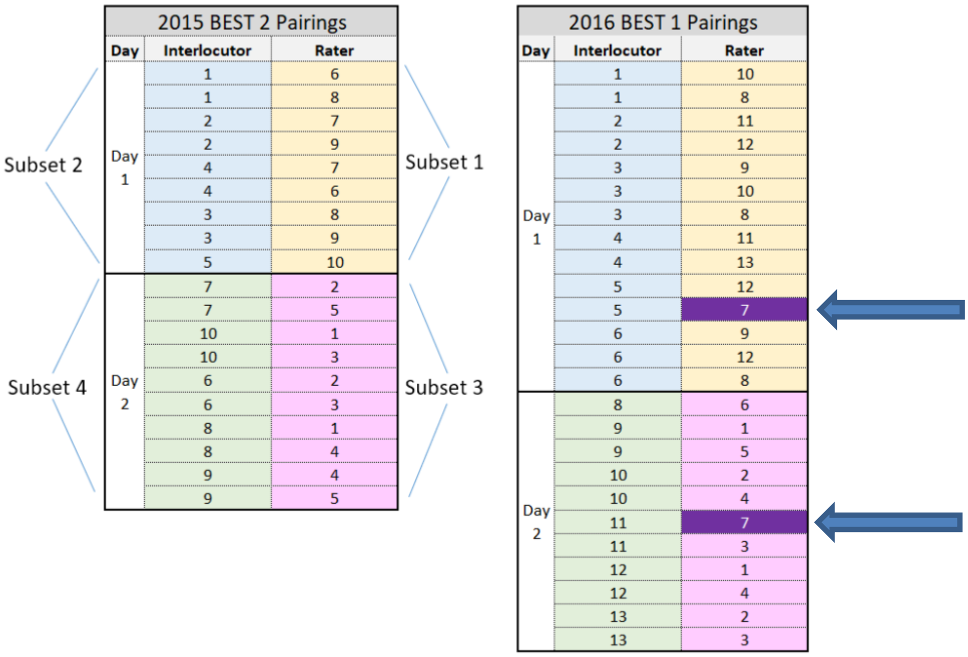
2015 BEST 2 and 4 subset summary

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



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

Subset 2
Subset 1
Subset 4
Subset 3

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:

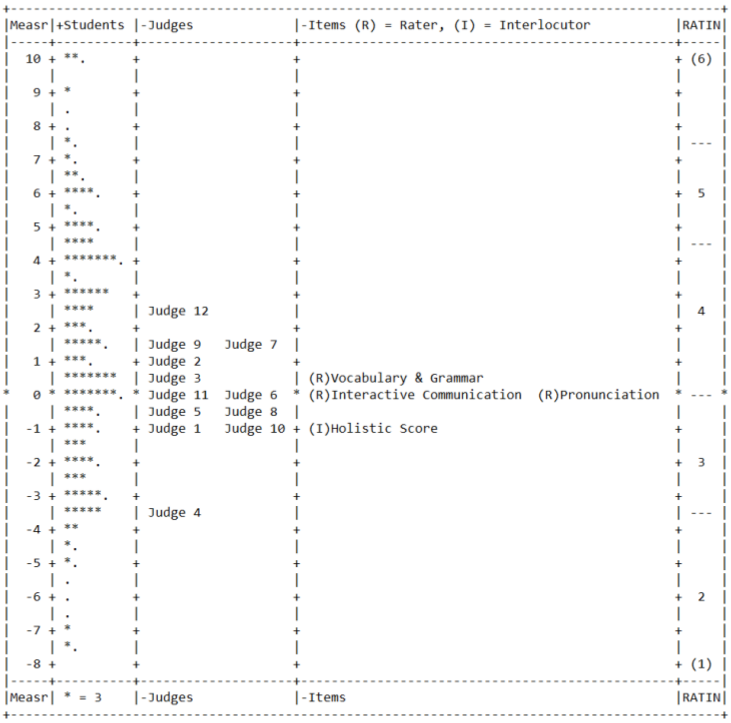
- Raters and interlocutors are not paired together more than once per test.
- To strengthen the integrity of the MFRM and provide as much data on teacher leniency and strictness as possible, teachers are separated after one test together (even if judging roles were to be reversed).
- Teachers do not 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.

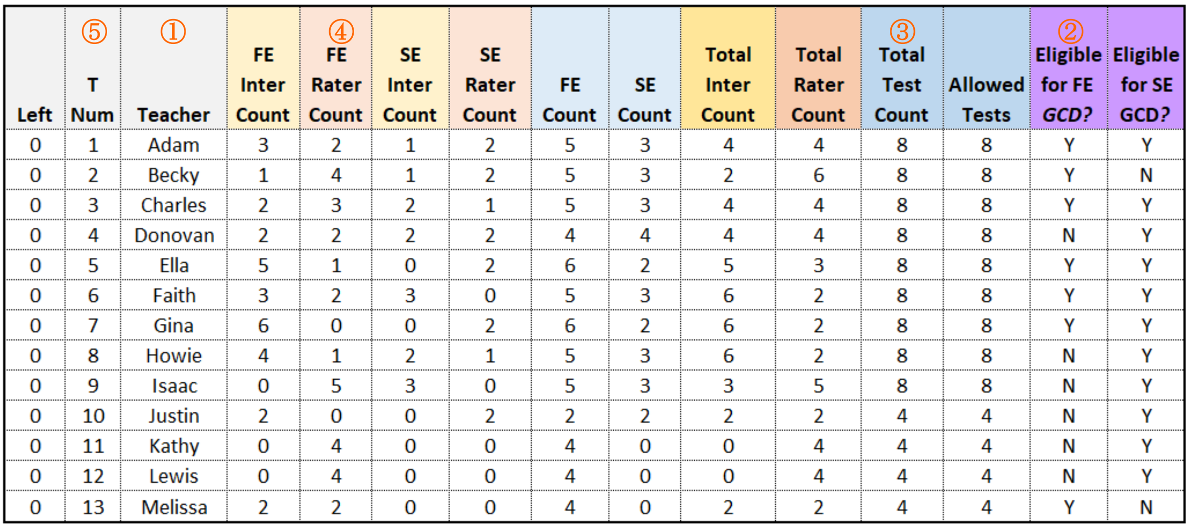
Lesson 2: Creating a scheduling and data entry database

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



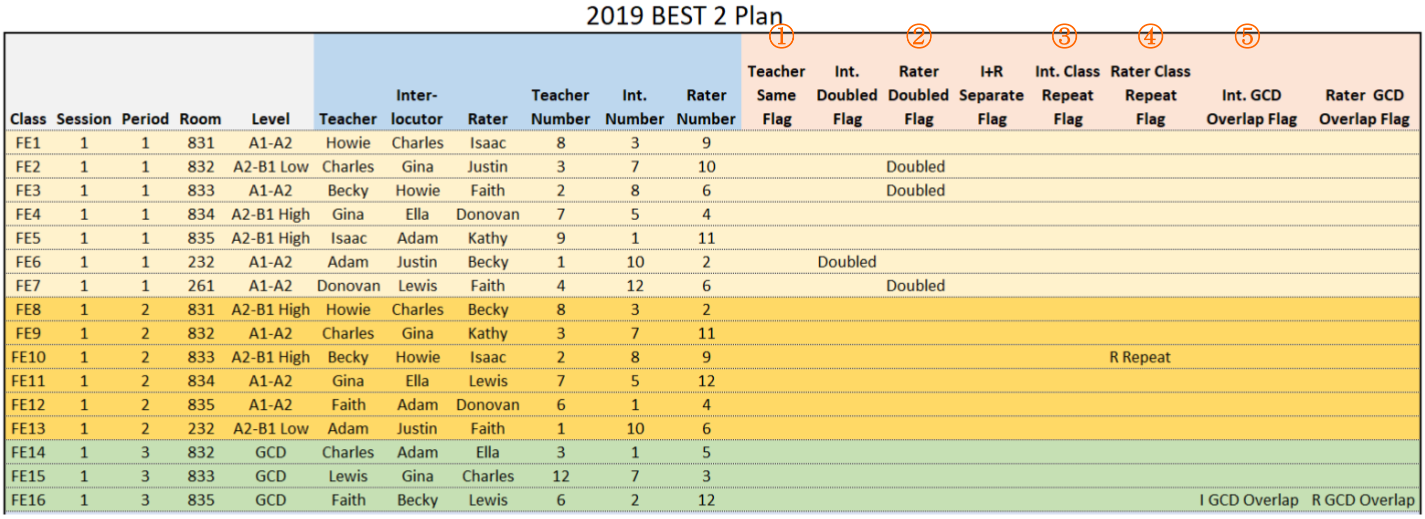
Left	⑤ T Num	① Teacher	FE Inter Count	④ FE Rater Count	SE Inter Count	SE Rater Count	FE Count	SE Count	Total Inter Count	Total Rater Count	③ Total Test Count	Allowed Tests	② Eligible for FE *GCD?*	Eligible for SE *GCD?*
0	1	Adam	3	2	1	2	5	3	4	4	8	8	Y	Y
0	2	Becky	1	4	1	2	5	3	2	6	8	8	Y	N
0	3	Charles	2	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.

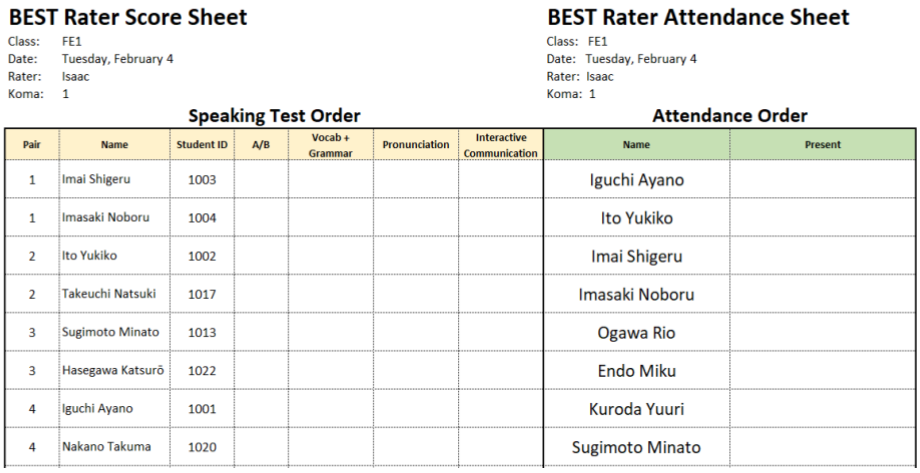
The 2019 BEST 2 and 4 plan as shown in Figures 4 and 5 above, contained 30 FE and 14 SE BEST sessions. In Figure 4, each of the 13 teachers was assigned to four or eight test sessions. Three teachers worked only as raters within FE, while two teachers were only interlocutors, and three other teachers 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

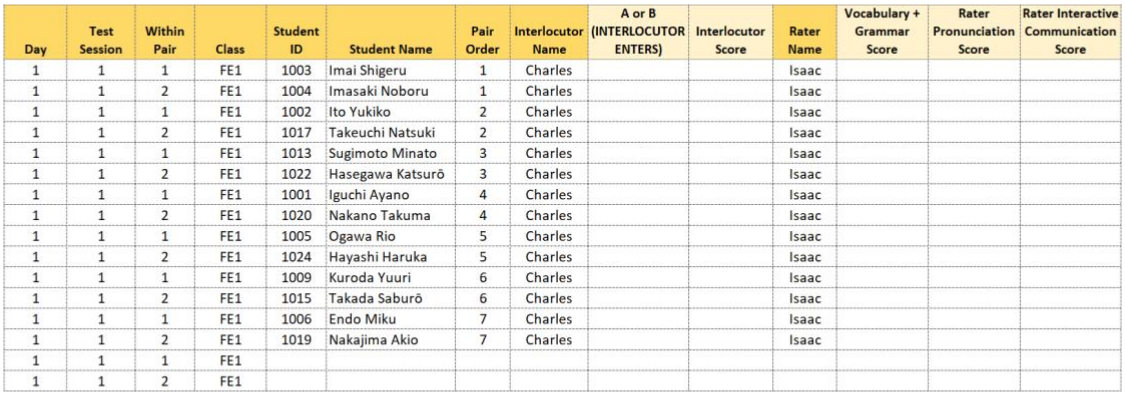
BEST Rater Attendance Sheet
Class: FE1
Date: Tuesday, February 4
Rater: Isaac
Koma: 1

Speaking Test Order							Attendance Order	
Pair	Name	Student ID	A/B	Vocab + Grammar	Pronunciation	Interactive Communication	Name	Present
1	Imai Shigeru	1003					Iguchi Ayano	
1	Imasaki Noboru	1004					Ito Yukiko	
2	Ito Yukiko	1002					Imai Shigeru	
2	Takeuchi Natsuki	1017					Imasaki Noboru	
3	Sugimoto Minato	1013					Ogawa Rio	
3	Hasegawa Katsurō	1022					Endo Miku	
4	Iguchi Ayano	1001					Kuroda Yuuri	
4	Nakano Takuma	1020					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										

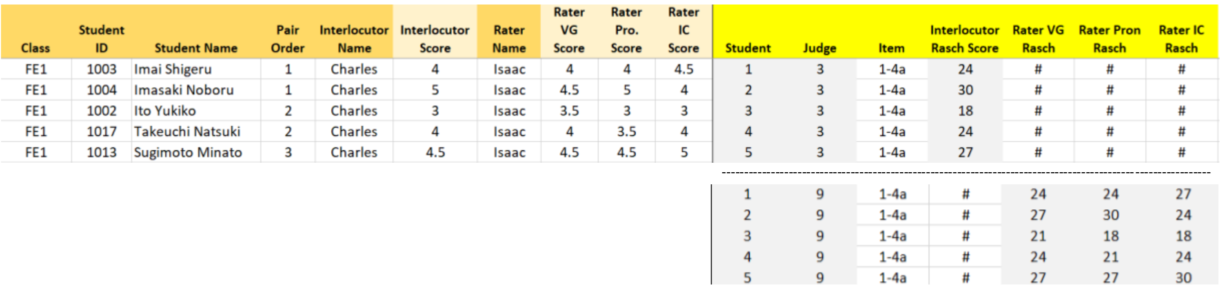
Sixteen rows (half the maximum class size) are stacked sequentially for each class and day, with rows without students intentionally kept blank: thus, a designated row for each student from Day 1, Student 1-A to Day 2, Student 8-B is assigned and only filled in if such student designation exists in each 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).

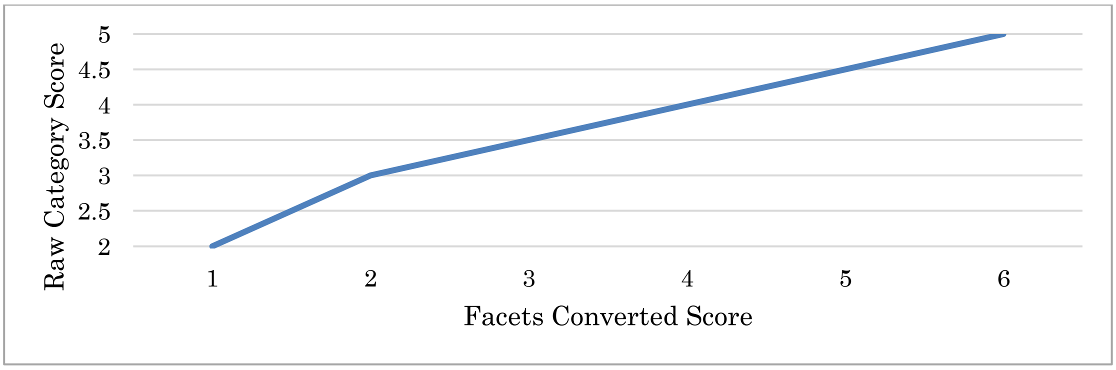
Table 6

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

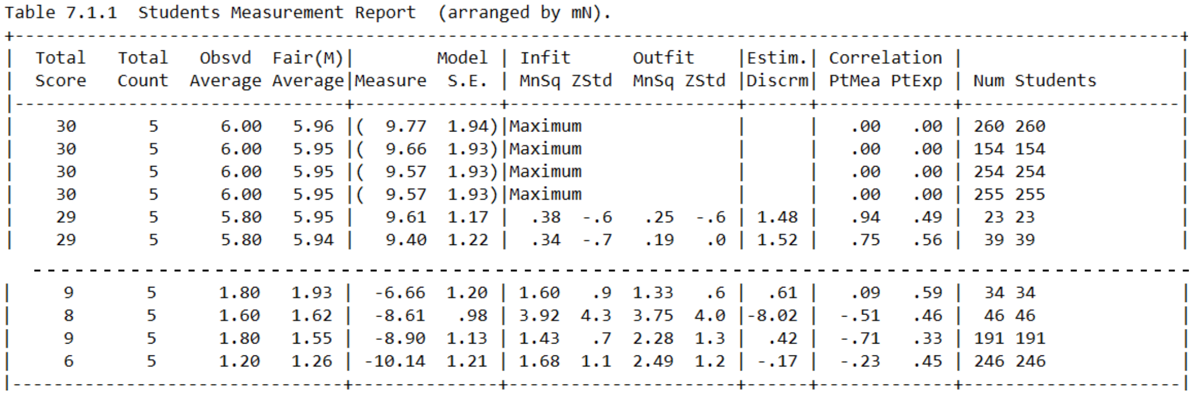
2016-18 BEST rating scale (raw and Facets converted scores)



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.

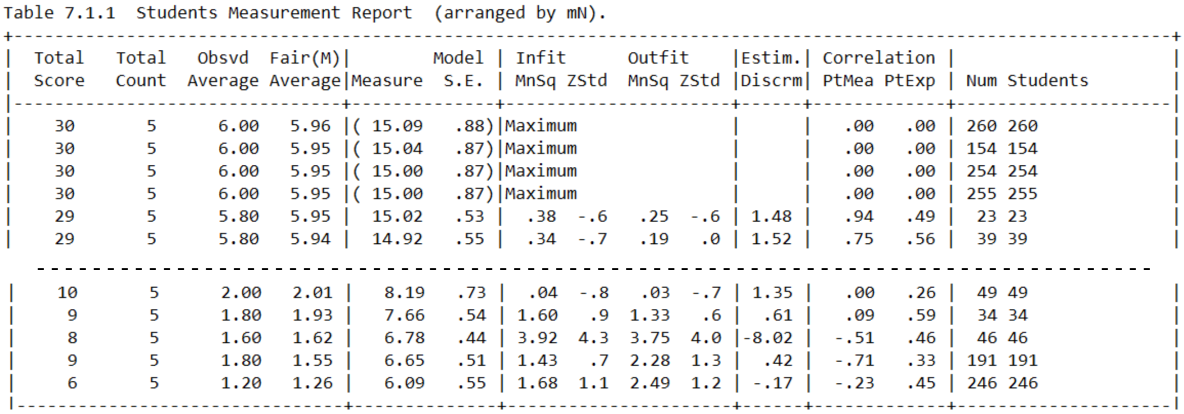
Table 7

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

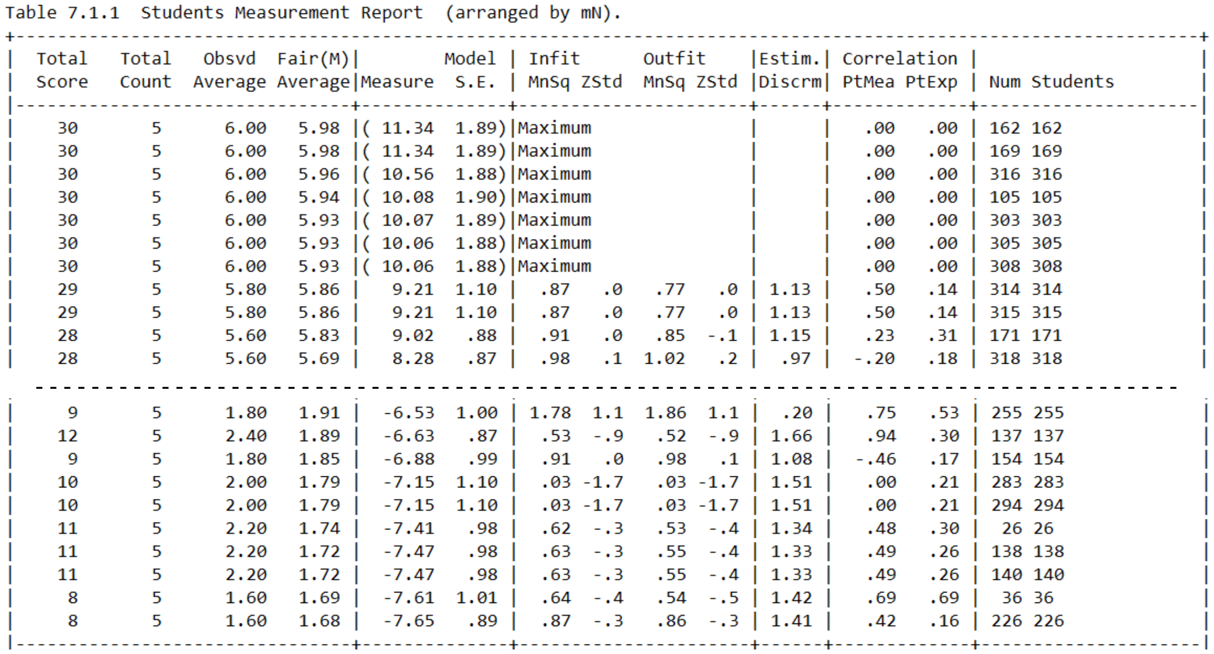


```
Table 7.1.1  Students Measurement Report  (arranged by mN).
+-------------------------------------------------------------------------------------------------+
| Total   Total   Obsvd  Fair(M)|        Model | Infit       Outfit     |Estim.| Correlation |                    |
| Score   Count  Average 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.

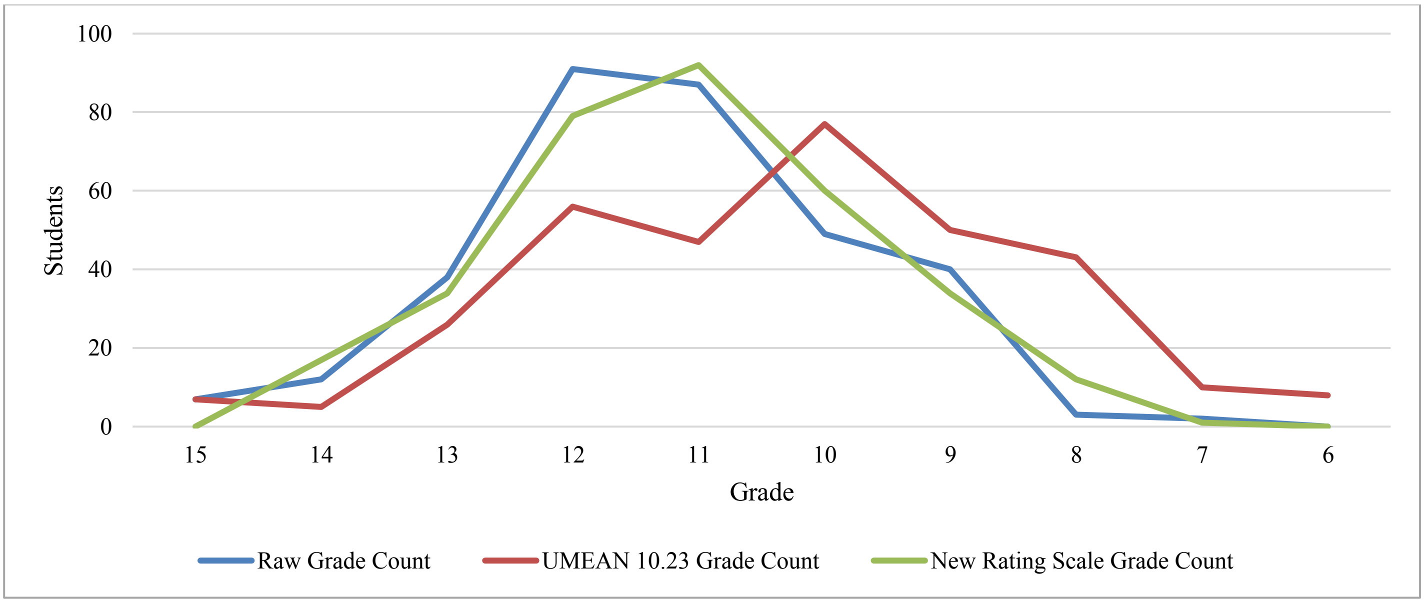
Table 8

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.*			

Call for Papers

Shiken: A Journal of Language Testing and Evaluation in Japan is seeking submissions for future issues on an ongoing basis. After checking the guidelines for publication below, please send your submission to the editor at tevalpublications@gmail.com.

Overview

Shiken aims to publish articles concerning language assessment issues relevant to assessment professionals, researchers, classroom practitioners and language program administrators. *Shiken* is dedicated to publishing articles on assessment in various educational contexts in and around Japan.

Article formats include research papers, replication studies, and review articles, all of which should typically not exceed 7000 words; as well as informed opinion pieces, technical advice articles, and interviews, all of which should typically not exceed 3000 words. Please contact the editor if you wish to submit an article that differs from these formats.

Novice researchers are encouraged to submit, but should aim for papers that focus on a single main issue. Please review recent issues of *Shiken* to understand the level of detail, depth of discussion and provision of empirical evidence that is required for acceptance.

Format

To facilitate double-blind peer review, the name(s) of the author(s) should not be mentioned in the manuscript. All citations and references to the author or co-author's work should read (Author, Year) e.g. "(Author, 2017)."

Articles should be formatted according to the guidelines of the *Publication Manual of the American Psychological Association (APA), 7th Edition*. Submissions should be formatted in Microsoft Word (.docx format) using 12-point Times New Roman font. The page size should be set to A4, with a 2.5 cm margin. The body text should be block justified, and single-spaced. Paragraphs should be separated by a blank line space. Each new section of the paper should have a section heading. Please review the most recent issues of *Shiken* for examples of section headings.

Research articles must be preceded by an abstract that succinctly summarizes the article content in less than 200 words. The reference section should begin on a new page immediately following the body text. Authors are responsible for comprehensive and accurate referencing including adding DOI or URL information wherever possible. Tables and figures should be numbered and titled. Separate sections for tables and figures should follow the references. Any appendices should be numbered and should follow the tables and figures.

Evaluation

All papers are double-blind peer-reviewed by two expert reviewers. Initial evaluation is usually completed within four weeks. The whole process from submission to publication is normally completed within six months.