CEFR Self-Assessment Listening vs. BULATS Listening: fall 2009

Antonio F. Smith

Introduction

The Common European Framework of Reference (2001) among other things, presents validated scales for measuring foreign language ability that are now in use throughout the European Union[1]. There, the scales serve as a kind of yardstick that can be used by learners, schools and employers regardless of European mother tongue and target language. Moreover this quasi "universality" of the Framework and its scales—not to mention the extraordinary investment spent in research and development—has made them attractive not only to institutions inside but also outside Europe. To assist such parties, the Council of Europe provides instructions for linking existing tests to the CEFR scales[2]. One recent example of a country outside Europe that has officially linked it's national language exams to CEFR is Taiwan, who's Ministry of Education officially commenced the linking procedure in 2005[3].

The inherent value in a common framework for language measurement can be compared to the inherent value of a transportation hub, used by airlines and logistics companies. Rather than directly delivering each person or parcel to a local address, which would require many vehicles and drivers, transport companies first collect

large numbers of people and packages at a hub, and then provide connections to outlying areas to meet individuals' requirements. Similarly, with the CEFR, those linking to it can easily make secondary linkages with the other participants, including such things as universities and employers. Taking up the case of Taiwan again, whereas very few people or institutions around the world clearly understand what Taiwan's GEPT (General English Proficiency Test) scores mean, a great and growing number of people and institutions around the world understand what the CEFR levels mean, so by linking GEPT to CEFR, Taiwan provides its language learners with vastly increased mobility internationally, and reciprocally, Taiwanese institutions can easily recognize the language qualifications of framework participants who want to use them inside Taiwan.

What is more, every new party that adopts the Framework contributes to the critical mass necessary for the CEFR to become a true global standard, the first and perhaps last of its kind. That is, the more countries that use it, the more useful the CEFR becomes, such that eventually there may be no good reason *not* to use it. Why would any country want to use a standard for foreign languages that few other countries understand, or no other country understands, when they can use one that virtually every other country understands? Can the main point of foreign language learning just be domestic use in the future, given the trend toward globalization and internationalization?

In view of such considerations and perhaps even more importantly the similarity of its multi-lingual situation to that of Europe, the former Osaka University of Foreign Studies and subsequently the School of Foreign Studies at Osaka University decided to adopt a CEFR-based achievement system (Majima 2007, Majima & Smith 2008). For many languages, this has involved creating tests linked to the CEFR, but for English some tests linked to CEFR already exist, including the Cambridge Main

Suite of tests, IELTS and BULATS. These last two are now administered by STEP in Japan and STEP has provided a free sitting of BULATS for the English program to study and report on.

In particular, the makers and administrators of BULATS are interested in knowing what its scores mean in terms of CEFR-based self-assessment, and the English Area Studies program at Osaka University is interested in knowing what CEFR-based self-assessment means in terms of objective tests linked to CEFR, such as BULATS. If self-assessed level and BULATS-assessed level, for example, are consistently close over time, then the validity and reliability of the two types of assessment corroborate each other. This in turn may create new possibilities, such as using self-assessment to sort incoming English minors into levels and using BULATS to prove achievement at the end of a term of study. As a part of the School of Foreign Studies, which has officially adopted CEFR, such a possibility is indeed interesting for the English program, and as the following results suggest, promising.

Method: CEFR-based self-assessment (hereafter, "SA")

Subjects: three first-year classes of English Area Studies majors, in fall 2009, 52 students

Instrument: WebCT/Blackboard version of Self-Assessment Checklists, Swiss Version (see note 1 for "Checklists" at Council of Europe website cited above), English-Japanese bilingual version.

Instructions: Students must complete each and every Checklist at every level

Scoring: 1 = "I can do this" (*dekiru*), 2 = "I can do this pretty well" (*daitai dekiru*), 3 = "I can do this a little" (*skoshi dekiru*), 4 = "I cannot do this" (*dekinai*). In the original instructions, if a student did not have full confidence in an ability s/he had to choose between "I can do this under normal circumstances" and "I can't do this". In

this forced choice situation, it is assumed that many Japanese will tend to choose "I can't do this" unless they have experienced that they definitely can. By introducing the three-option system, Japanese are more likely to affirm intermediate levels of confidence when they apply.

Interpretation of scores: 1 = 100%, 2 = 100%, 3 = 0, 4 = 0. A student qualifies for a level if the majority of possible items are affirmed with 1's or 2's.

The levels are given the following numerical representation in this study (1 higher than ALTE at each level): A1=1, A2=2, B1=3, B2=4, C1=5, C2=6.

BULATS

The same set of students took BULATS, listening and reading only, in house, administered by STEP in spring 2010.

Scoring: 1-19 = A1, 20-39 = A2, 40-59 = B1, 60-74 = B2, 75-89 = C1, 90-100 = C2

The levels are given the following numerical representation in this study (1 higher than ALTE at each level): A1=1, A2=2, B1=3, B2=4, C1=5, C2=6.

DATA

The following table shows the results of SA beside the results of BULATS by student; 47 students returned data for both assessments. However, the data of the two students scoring only 50% in A1 (2/4) and higher in subsequent checklists are excluded because they do not fit the pattern of other self-assessors or the intended progression of the scales (One of these students scored 60 on BULATS, the other 62; their data is shown with a single horizontal slash for the reader to examine; both are male). Therefore, the data of a total of 45 students are considered. Details used for interpreting the data follow.

For B2, 6 is the maximum number of points a student can receive (6 Max), and while for other levels a majority is required to pass, which in the case of B2 would be 4, for B2, if the student receives only 3 points, he can pass if the items not counted for regular points are all 3's (i.e., "*sukoshi dekiru*") and not 4's (4 = "*dekinai*"); this is introduced to reduce underestimation in this range and does not apply elsewhere.

Grey colored cell = pass/qualifies for that level

Vertical lines = over estimator in SA (i.e., author interprets that these students affirm a level higher than that affirmed by their BULATS score).

Horizontal lines = under estimator in SA (i.e., author interprets that these students affirm only up to a level lower than that indicated by their BULATS score).

Multiple diagonal lines = please notice; in C1, two students got three points, which is not enough to pass according to the author's rules, but one of them, BULATS 68, went on to fully affirm C2; interestingly; this student is from Singapore (mother tongue Chinese) where English is widely spoken, which may have given her the confidence to affirm C2.

Analysis of listening results:

SA A1 Affirmed 45/45

Under estimators: 0 (given exclusion)

Over estimators: NA

SA A2 Affirmed 45/45

Under estimators: 0

Over estimators: 0

BULATS scores indicate every student reaches A2 level

SA B1 Affirmed 42/45

Under estimators: 3 (Female: 2; Male: 1)

All but three students affirmed B1. Given their BULATS scores, these three

Table 1: Listening, SA data and interpretation, plus BULATS

St	A2 6M	B 1a	B 1b	B 1c	B 1d	B 1e	B 1f	B1 6M	B 2a	B 2b	B 2c	B 2d	B 2e	B 2f	B2 6M	C 1a	C 1b	C 1c	C 1d	C 1e	C 1f	C1 6M	C 2a	C1 1M	B-L
1F	6	2	2	2	2	2	3	5	3	3	3	3	3	4	0	4	4	4	4	4	4	0	4	0	39
2F	6	1	2	2	1	1	4	5	4	3	3	3	4	4	0	4	3	4	4	4	4	0	4	0	41
3M	6	1	1	2	1	2	2	6	3	1	2	2	2	1	3	4	3	4	3	2	4	1	4	0	45
4F	6	2	2	2	1	2	2	6	3	3	3	3	3	2	1	3	4	4	3	3	4	0	3	0	45
5M	6	1	1	1	1	1	2	6	2	1	3	3	3	2	3	4	3	2	2	3	4	2	4	0	47
6F	5	3	3	4	3	3	4	0	4	4	4	4	4	4	0	3	4	4	4	4	4	0	4	0	47
7F	6	2	2	2	2	2	2	6	3	3	3	3	3	3	0	3	3	3	2	3	3	1	3	0	54
8F	6	1	1	1	1	2	2	6	3	3	3	4	4	3	0	4	4	3	3	3	4	0	4	0	54
9M	6	1	1	1	1	1	2	6	2	2	2	2	3	3	4	2	3	2	3	3	3	2	3	0	56
10M	6	1	1	1	1	1	1	6	2	2	2	3	1	1	5	4	4	3	3	4	4	0	4	0	56
11M	6	1	1	1	1	1	2	6	2	2	2	2	3	2	5	3	3	3	3	3	4	0	4	0	56
12F	6	1	1	2	1	1	2	6	3	2	3	3	4	3	1	4	4	4	3	4	4	0	4	0	56
13F	6	1	1	2	2	2	2	6	2	3	3	3	3	3	1	3	3	4	4	4	4	0	4	0	56
14F	6	1	2	2	2	3	3	4	3	3	3	3	4	4	0	3	4	4	4	4	4	0	4	0	56
15M	6	2	3	3	3	2	3	2	4	3	3	3	4	4	0	3	4	4	3	4	4	0	3	0	56
16M	6	1	1	1	1	1	2	6	2	2	2	2	2	2	3	1	2	2	1	3	3	4	3	0	57
17F	6	1	1	1	1	1	1	6	1	2	3	3	3	3	2	4	4	3	3	4	4	0	4	0	57
18M	6	1	2	2	1	1	2	6	2	1	3	3	3	3	2	4	3	3	2	3	4	1	4	0	57
19F	6	2	3	3	3	3	3	0	3	3	3	4	4	4	0	4	4	4	4	4	4	0	4	0	57
20F	6	1	1	1	1	1	1	6	2	1	1	2	3	3	4	3	3	3	3	3	4	0	4	0	58
21F	6	1	1	1	1	1	1	6	2	1	2	2	3	4	4	4	3	3	3	3	4	0	4	0	58
22M	6	1	2	2	1	1	2	6	2	2	2	3	2	2	5	3	3	3	3	4	3	0	4	0	58
~~23M~~	~~4~~	~~3~~	~~2~~	~~3~~	~~2~~	~~2~~	~~3~~	~~3~~	~~2~~	~~3~~	~~2~~	~~2~~	~~1~~	~~4~~	~~4~~	~~4~~	~~1~~	~~2~~	~~3~~	~~2~~	~~2~~	~~4~~	~~2~~	~~1~~	~~60~~
24M	6	2	2	2	2	2	2	6	3	3	3	4	4	4	0	3	4	3	3	3	3	0	3	0	60
25F	5	1	2	3	3	2	2	4	3	3	3	3	3	3	0	2	3	4	3	3	4	1	3	0	60
26M	6	2	2	2	2	2	3	5	3	3	3	3	3	3	0	3	3	3	3	3	3	0	4	0	60
27M	6	1	1	1	1	1	1	6	1	1	2	2	2	1	5	3	3	3	4	3	4	0	4	0	62
28M	6	1	1	1	1	2	2	6	3	2	2	2	2	3	4	3	3	4	3	3	4	0	4	0	62
29M	6	1	1	1	1	1	1	6	3	3	3	4	4	3	0	4	4	4	4	4	4	0	4	0	62
30M	5	2	2	1	3	2	3	4	2	3	3	3	3	3	1	3	3	4	4	4	4	0	4	0	62
~~31M~~	~~3~~	~~2~~	~~2~~	~~3~~	~~2~~	~~3~~	~~2~~	~~4~~	~~3~~	~~3~~	~~3~~	~~2~~	~~2~~	~~2~~	~~3~~	~~3~~	~~3~~	~~2~~	~~3~~	~~3~~	~~3~~	~~1~~	~~2~~	~~1~~	~~62~~
32F	6	1	2	1	2	2	2	6	3	2	2	2	3	3	3	3	2	3	3	3	3	1	3	0	64
33F	6	1	2	1	1	2	2	6	1	2	2	2	3	2	5	3	3	2	3	3	3	1	3	0	67
34M	6	1	1	1	1	1	1	6	2	2	2	2	3	2	5	3	3	3	3	3	3	0	4	0	67
35F	6	1	1	1	2	1	2	6	2	1	1	2	2	2	6	3	1	3	3	2	2	3	2	3	68
36F	6	1	1	1	1	1	2	6	1	3	2	2	3	3	3	2	3	2	3	3	2	3	3	0	68
37M	6	1	1	1	1	1	1	6	1	1	2	3	3	3	3	4	4	3	3	4	4	0	4	0	70
38F	6	1	2	2	2	2	3	5	3	4	4	4	4	3	0	3	4	3	4	4	4	0	4	0	70
39F	6	2	2	3	2	2	3	4	4	4	4	4	4	3	0	4	4	3	4	4	4	0	4	0	70
40F	6	1	1	1	1	2	2	6	2	3	3	4	4	4	0	4	4	4	4	4	4	0	4	0	72
41F	6	1	2	1	1	1	2	6	2	3	3	3	4	4	1	4	4	4	4	4	4	0	4	0	72
42M	6	2	2	2	2	2	2	6	3	2	3	4	3	3	1	4	4	4	3	4	4	0	4	0	72
43M	6	1	1	1	1	1	1	6	1	1	1	1	1	1	6	2	2	2	1	2	3	5	3	0	79
44F	6	1	1	1	1	1	1	6	1	1	1	1	1	1	6	1	1	1	1	1	1	6	1	1	84
45F	6	1	1	1	1	1	1	6	2	2	2	2	1	2	6	2	3	2	2	2	3	4	4	0	86
46F	6	1	1	1	1	1	1	6	1	1	1	1	1	1	6	1	1	1	1	1	1	6	1	1	92
47F	6	1	1	1	1	1	2	6	1	2	1	2	2	1	6	2	2	1	2	2	1	6	2	1	93

underestimate their listening ability. Speaking to students in small groups as their teacher leads the author to concur with BULATS; no student seemed unable to understand the author when he spoke simply and carefully, even when discussing newspaper articles and the like. Changing the value of "*sukoshi dekiru*" to 0.33 brings only one of the three up to this level, so it is not a useful change. A teacher or program wishing to set classes by level could check such lowest self-assessors by a mini listening test and/or an interview testing B1 can do's.

Over estimators: 1 (short by 1 point)

SA B2: Affirmed: 23/45

Under estimators: 10/18 (Female: 5, Male 5), high B2 and low B2

Just qualified students in this area, BULATS 60-62, tend to lack confidence about their level; five out of seven in this range underestimate and all but one of the Five are male; therefore, it seems that being just qualified B2 trumps gender (to be discussed below). In the middle, BULATS 64-68, five out of five affirm in alignment with BULATS. Then, surprisingly, at the upper end of the B2 range (70-72), five out of six underestimate (and of the five who underestimate in this range, four are female). It is a bizarre phenomenon to see people in the 70-72 range underestimating when they are in fact the strongest B2's in this study. Perhaps it is not until this level that students become acutely aware of how far they must go to reach the advanced level displayed by a few of their classmates. Further iterations of this study will show if this is an anomaly or a real trend.

Gender: Like the lowest B2 males (60-62), the highest B2 females (70-72) tend to under estimate; perhaps these females need to be counseled not to compare themselves with the C1-C2 level students, who are mainly female (4/5) and returnees. Their behavior is in stark opposition to that of the low B1 males (45, and 47) who overestimate their level to be B2, and that of males in the group of seven high B1 students at BULATS 56-57 in which 4/5 males overestimate and no female overestimates. There seems to be a gender phenomenon in Japanese culture where

it is not uncommon to find lower level young men who have learned to display confidence about their abilities even to the point of overstating them, and find higher level men who are not yet confident enough to claim the level they have just barely achieved. Conversely, the second highest tier of females (and one male) at very high B2 seemed to have learned to doubt their abilities even when they are well-developed. In fact, these are the highest tier among non-returnees and what may have allowed them to climb so high is their belief that they have not yet studied enough and must always study more. The very top level students on the other hand can recognize their ability with confidence.

Over estimators: 9/45 (M 7 /F 2), **6/45 (M 6/F 0)** or 2/45 (M2/F0)

Nine students overestimated their ability according to BULATS' cut score of 60. However, given students' unfamiliarity with the test, business English, and perhaps British English, it may be most accurate to assume that the three students with scores of 58, almost immediately below the cut score of 60, do in fact have B2 level. If so, the four "over estimators" scoring 56, 56, 56 and 57 are really on the border between B1 and B2 and would not be grossly misplaced if they joined a class targeting B2 level, although they would be at the bottom of the class (similarly, students with 56 who did not affirm B2 could be all right at the top of a B1 listening class, but if a school allows it, it might be advisable for such students to try to test up a level at mid-term). The two students with 45 and 47, however, have grossly overestimated and would stand out in a B2 listening class, so the teacher could advise them to move down a level.

Gender: All of the over estimators in the range of BULATS 45-57 are male, yet males make up only 18 of the 45 subjects.

Accurate estimators: Every student with a score above 72 affirmed B2 level.

SA C1 Affirmed by 8/45

Underestimators: (assess as C1 but are C2 according to BULATS) 0

Overestimators: 1/6 (male)

One male student with a BULATS score of 57 just qualified for C1 in SA using the author's rules (4 out of 6). Either he simply overestimated or for him "*daitai dekiru*" is in fact similar to the meaning others ascribe to "*sukoshi dekiru*". This student has risen to very nearly the top of B1 and is feeling perhaps a bit too sure of himself.

Every student scoring 79 or higher affirmed C1

C2:

Underestimators: NA

Overestimators: 2/4, 1/4 or 0/4

One student with a BULATS score of 84 self-assessed as C2, using the authors' rules, yet she may not really be an over estimator. This student spent a few years at an American high school. It is possible that her score was lower than it should have been because of the British bias of the test; however, it could also be due to other factors, such as unfamiliarity with the test in general, the fact that its score does not affect her grade, or that some C2 academic skills (CALPS) were not fully developed at her high school, although her basic language skills (BICS) seem native-like. On the other hand, maybe as a nearly borderline case, she just overestimated. Another student, with a BULATS score of 68, also marked C2 as "*dekiru*", but she is from Singapore, and maybe really is at C2 level when listening to the variety of English spoken there, as opposed to the variety(ies) used on BULATS.

A probable problem with the C2 checklist is that it consists of one item, so any mistake in interpreting it can cause a bad self-assessment.

Results

All in all, there is a close match between SA listening and BULATS listening for this group of students in most levels. However, there is some variability in the border area between B1 and B2, in the range of BULATS 56-62, which also happens to be where the majority of students scored. This discrepancy would disappear however, if

the cut score were 59+/-3, rather than simply 60. Moreover, given the unfamiliarity of the test, the British English bias and the fact that students know the test score does not affect their grade, it should probably be expected that students underperform a little, so 59 may be a better cut score for them than 60.

It should also be expected that students very near the border waver around the cut score a little because SA is imperfect and subjective, with each person varying at least slightly in how s/he interprets the questions and his/her own abilities. The large number of apparent mismatches in this range does not therefore impugn BULATS's ability to discriminate well; in fact it probably confirms BULATS' discrimination ability because the mismatches occur within +/-3 of BULATS's cut of 60, minus one, 59.

If future studies confirm this level of accuracy, SA could be relied upon to sort students into preliminary levels and teachers could send over and under estimators up or down as needed after witnessing their performance in class or by other means, such as a mini-test for questionable cases.

Nevertheless, it may be possible to improve the consistency of SA by the following:

1. More instruction or training for students to better understand the items.
2. More questions; the new version of S.A., that used by CERCLEs, will have more items in each level, which will provide more data and thus more certainty about levels.
3. Checking if any items stand out as being frequently affirmed out of sequence and rewording if needed.

References

Council for Cultural Co-operation, Education Committess, Modern Language Division, Stasbourg (2001) Common European Framework of Reference for Language: Learning, teaching, assessment. Cambridge: Cambridge University Press

Majima, Junko 2007.「言語教育における到達度評価制度に向けて—CEFRを利用した大阪外国語大学の試み—」『間谷論集』創刊号　日本語日本文化教育研究会 pp. 3-27

Majima, Junko. & Smith, Antonio. F. (2008) *The Educational Impact of the CEFR on a Japanese National University* Paper presented at the 3rd international meeting of ALTE (The Association of Language Testers in Europe), Cambridge University

Notes

1) http://wpedia.goo.ne.jp/enwiki/CEFR and http://www.coe.int/T/DG4/Portfolio/?M=/main_pages/levels.html

2) http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

3) http://www.lttc.ntu.edu.tw/academics/geptresearch/Abstract%20-%20Using%20the%20CEFR%20in%20Taiwan.pdfcitation http://www.ealta.eu.org/conference/2007/docs/pres_sunday/Wu&Wu.pdf