CEFR, Self-Assessment, TOEIC and BULATS

Antonio F. Smith

1. Introduction

In 2005, Osaka University of Foreign Studies implemented a CEFR-based achievement system for all of its languages; then after the merger with Osaka University, the School of Foreign Studies also officially adopted the CEFR-based achievement system (Majima 2007, Majima & Smith 2008). Under such a system, it is important to understand in detail what the CEFR scales mean via the main CEFR text[1] and critical reviews[2], and then use that knowledge to accurately judge the starting level of students, set reasonable teaching/learning targets, choose appropriate materials/make appropriate curriculum and set minimum achievement levels.

Traditionally, English Area Studies majors have been required to reach TOEIC benchmarks, (English majors: 680 1st year, 730 2nd; English sub-majors 550 by end of 1st year to continue to 2nd) but TOEIC scores alone did not provide sufficient detail about CEFR level and particular *can-do's*. Moreover, although Educational Testing Services (hereafter, ETS), makers of TOEIC and TOEFL have published equivalencies between their tests and the CEFR, their published cut-scores have changed substantially over time in terms of C1 cuts[3], and their tests are not designed *a priori* to measure the CEFR constructs, so it was difficult to be sure

which version of the cut-scores, if any, should be relied upon.

Of course, the surest way to measure students' levels in terms of CEFR would have been to use the Cambridge main suite of tests, IELTS or BULATS as they are designed to measure CEFR level or are linked to it by Cambridge ESOL, the undisputed authority regarding English tests linked to CEFR. However, for practical reasons, such as cost and infrastructure, that was not possible at the time. Dialangue also had demerits that made requiring it of every student impractical.

Therefore, in 2007 the author decided to ask 2nd year English majors to complete the Swiss version of the Self-Assessment Checklists from the European Language Portfolio [4] (see Appendix B) to provide subjective data to compare with TOEIC, help estimate students' levels, inform curriculum development/teaching targets and contribute to the formulation of realistic achievement goals (Majima & Smith 2008, Smith 2009). Results showed that students were, on average, achieving or exceeding the achievement the goal, B2+, but were low B2 in terms of speaking and listening.

As a result, in the following year, first year students were given an "extensive listening" assignment outside of class, and inside class, student-talk-time based on the extensive listening assignment was maximized. 1st year students also submitted a hard copy of the Self-Assessment Checklists at the beginning of the 1st semester, April 2008, revealing their perceived overall level to be B1. Therefore, B2 was confirmed as the "teaching/learning target" for 1st year native-speaker teacher classes. Students then completed a bilingual WebCT version of the Self-Assessment Checklists at the end of the 2nd Semester, January 2009. Students also took TOEIC (listening and reading version) at the end of the semester, if they were not already in possession of a satisfactory score. Then, in April of 2009, thanks to assistance from STEP, the same group of students took STEP-BULATS (listening and reading version). This paper compares and analyses the results of all of the above-mentioned assessments.

2. Reconciling ETS, BULATS and self-assessment data with teacher evaluation

2.1 ETS's most recent standard-setting cut-scores are likely too high

2.1.1 The C1 cut is too high to be useful in Japan

The most recent reference document for ETS cut-scores is: TOEFL iBT Research Report, TOEFL iBT-06, June 2008, Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology, Richard J. Tannenbaum and E. Caroline Wylie[5].

The cut-scores that resulted from this study were high, especially for the C levels—B2: 785; C1: Listening 495, Reading NA; Speaking 200/200, Writing 200/200; C2: NA. Probably, the C1 cut is too high to be of practical use by TOEIC users in Asia. In Japan, for example, which relies heavily on the TOEIC, almost no one gets a perfect or nearly perfect TOEIC score, so, with the above cuts, the test does not allow employers and schools to discriminate between B2 (upper intermediate) and C1 (advanced/operational proficiency). Something must be wrong.

2.1.2 ETS 2008 Standard-Setting cut score for C1 significantly higher than 2005

As mentioned earlier, ETS has published different cut-scores over time. For example, in a 2005 standard-setting study, by Tannenbaum and Wylie, the following scores in Table 1 were published[6].

Table 1 ETS cut scores 2005

	B1	C1
[Paper-based] TOEFL	457	560
TOEIC	550	880

Why such a big difference between this standard setting session and 2006, in terms of C1? Part of the answer may have to do with this,

> "The variability (standard deviation) of the panelists' judgments for the B1 level decreased from round one to round two, indicating a greater degree of panelist consensus; the variability increased somewhat for the C1 level between the two rounds[7]."

In the end, standard-setting results are the distillation of a collection of educated guesses by experts whose views are not identical, so it is not surprising that the two sessions had different results regarding C1. However, the magnitude of the difference, about 100 points, suggests that the results of one or both of the standard-setting sessions may not be right, and that the 2008 C1 cut could well be too high.

2.1.3 TOEIC-TOEFL chart by ETS Canada: 637 TOEFL(110 iBT)/C1= about 957 TOEIC

ETS Canada provides the following correspondence between TOEIC and TOEFL[8]

Table 2 Rough correspondences between TOEIC and TOEFL (listening and reading)[9]

TOEIC	TOEFL
950-990	628-677
900-950	608-628
850-900	588-608
800-850	569-588
750-800	549-569
700-750	529-549
650-700	509-529

This page hedges about the degree to which TOEIC and TOEFL scores can be compared given that the two tests do not measure exactly the same set of constructs, and declares itself to be a rough guide. Nevertheless, the discrepancy between this and the 2006 standard-setting study in terms of C1 is significant. That is, the 2006 study, published in June 2008, declares that a TOEFL score of 637 paper (110iBT) amounts to roughly the entry of C1. However, 637 falls within the 950-990 TOEIC range above. To estimate a more precise number, for TOEFL subtract 628 from 677, which equals 49, and divide by 10, for units of 4.9. For TOEIC, subtract 950 from 990, which equals 40, and divide by ten. This gives a rough correspondence for the top-end of these two tests, based on the ETS Canada table.

Table 3 TOEIC-TOEFL Correspondences based on Table 2

TOEIC	TOEFL
990	677
986	672.1
982	667.2
978	662.3
974	657.4
970	652.5
966	647.6
962	642.7
958	**637.9**
954	632.9
950	628

This then suggests a 637.9 TOEFL is equal to about 958 TOEIC. Therefore, a little less than 958 TOEIC, perhaps, 957, may be about equal to 637 TOEFL. This is still a very high score, but much less than the nearly perfect TOEIC score of the 2006 ETS study published in 2008.

2.1.4 U.S. universities' English requirements reveal their concept of C1 lower than that of ETS

In most cases, American universities accept foreign students based on either TOEFL or IELTS, and presumably they select minimum requirements for each test that produce students with roughly the same English proficiency. Now, thousands of American universities accept IELTS, and most require a minimum band score of 6.5-7, which is Cambridge ESOL's cut-score for C1 on that test[10]; in a few cases American universities require as low as a 6 and as high as a 7.5.

On the other hand, American universities tend to require 70-100 TOEFL iBT. The University of California campuses usually take 90 for undergraduates, except Berkeley, which requires 100. There seems not to be a single university that requires TOEFL 110 iBT, 637 paper, for undergraduates.

In sum, based on their IELTS requirements, it can be deduced that universities want students with C1 status (or close to it); however, based on their TOEFL requirements, it can be seen that they, in effect, equate C1 to being *at least* 10 points lower than the ETS cut-score of 110. From this, we can conclude that both Cambridge ESOL and thousands of American universities view the threshold for C1 substantially lower than do ETS.

Given sufficient reason to doubt ETS' most recent C1 cut-score, what might have gone wrong?

2.2 ETS's Standard-Setting Method

In the study, 22 expert participants from around Europe first studied the sets of skills/abilities that define entry into the various CEFR levels. Then they were asked to judge the minimum TOEIC score that would be required to guarantee that an examinee possessed the skills/abilities of a particular CEFR level. Finally, they discussed their results and at least 65% of judges had to agree on the cut-score for a given CEFR level in order for it to be published.

2.2.1 ETS's Standard-Setting Method vs. Self-Assessment Checklist Method

ETS's procedures for standard setting seem good, so why/how could there be any problem? One possibility is suggested by the difference between the instructions given by North to those using the Self-assessment Checklists and those given by ETS to the standard-setting panelists.

First of all, for context, I include the following quotation from the Self-assessment checklists from the Swiss version of the European Language Portfolio to show that the Checklists are closely related to CEFR and that they can be used to identify proficiency in terms of the CEFR scales for various purposes, including those of the author.

> These checklists are based on the common reference levels elaborated in the *Common European Framework*; they are thus closely related to the illustrative scales set out in Appendix 1. The Swiss ELP explains that the checklists can be used in two ways: (i) to review one's overall proficiency in a particular language prior to updating one's language passport at the beginning or end of an extended period of learning; and (ii) to monitor one's learning progress, perhaps in relation to a particular skill or skills. Like the illustrative scales, the checklists can also be used to plan a course of learning and to identify appropriate learning tasks.

However, there is another, equally important quotation from the Checklists: "If you have over 80% of the points ticked, you have probably reached Level C1". This notice appears on every page of the Self-Assessment Checklists. In fact, every CEFR level checklist in every skill area has an identical statement, except for the CEFR level at the end of the sentence. This caveat seems legitimate for two reasons: 1. If a person can truthfully affirm all of the B2 can-do's and 80% of the C1 can do's, for example, s/he is much more at C1 level than at B2. 2. Each section

of the Checklists include numerous can do's, and learners are not identical in what they have studied or in their confidence level, so it is reasonable to expect that not everyone who has just entered a level will affirm exactly the same skills. Given that this 80% rule is important enough to appear on every section of every page of the Checklists written by North, the highest authority, then should it not also be used when mapping TOEIC to CEFR?

In terms of standard-setting procedure, maybe not. For the people participating in standard setting, it would be very difficult to factor in the 80% rule of the Checklists in a statistically valid way: Which descriptor(s) does not necessarily need to be affirmed, in the context of which other ones being affirmed etc? There would be such a multiplicity of scenarios that it would be virtually impossible for participants to agree. Therefore, it is not surprising that the job of participants in the 2008 study (as far as can be discerned) was to select the minimum TOEIC score required to affirm 100% of the "just qualified" level descriptors used. These descriptors are very similar to those in the Checklists (see appendix A for level descriptors used in Standard-Setting and B for a sample of can-do's from the Checklists). Nevertheless, if possible, something should be done, in a principled way, to produce more reasonable cut-scores.

2.2.2 ETS Globals' revised C1 cut score

After the 2006 standard-setting study, published in 2008 (Tannenbaum and Wylie), ETS Global issue guidance that revises the C1 cut-score for reading from NA to 455[11]. It was published in 2008 by ETS global with this explanation:

> At least two-thirds of the panel concluded that the TOEIC® Listening, Speaking and Writing sections ranged from the A1 level to the C1 level, and that the Reading section ranged from the A1 level to the B2 level. The Reading section did not meet the two-thirds criterion at the C1 level; 45% of the panelists (10

of 22) recommended a cut score at this level. Although the two thirds criterion was not satisfied, ETS understands that decision makers may still need to have a reference for what a potential TOEIC® Reading cut score may be at this level. It is with this awareness that the C1 value of 455 is reported.

Other ETS sites include the same cut[12]. However, one must wonder why 12 of the 22 panelists did not see evidence in TOEIC of C1 reading. It may have been the case that the descriptors used for this section of the standard-setting study lacked sufficient detail for some panelists to decide whether or not C1 reading could be proved by TOEIC. See descriptors used below.

Reading skills of just-qualified C1

Needs to re-read; more effort required than C2 for complex, extended text in all fields of interest.

Identifies or infers opinion, intention, feelings of writer.

This is far less detailed than the descriptors for other sections (see Appendix A).

Nevertheless 455 reading plus 490 listening make a combined score of 945, which would include many more advanced Japanese test-takers. However, 945 would still exclude many students in this study with Self-assessment, BULATS assessment and teacher assessment of C1. What could be done in principle to bring the cut score down to a more realistic level?

2.2.3 "Smith-mapping": post hoc application of 80% rule

The simplest solution is simply to apply North's 80% rule from the Checklists to the ETS cut scores. The ETS cut-scores represent 100% affirmation of the set of the abilities/skills required to enter a level, so 80% of that result should be about right. To approximate 80%, one can take the difference between the cut-scores for

two lereis and divide by five, producing units of 20%. 80% above the lower cut can then count, probably, as qualifying for a level. For example, using the C1 cut-score of 945 minus the B2 cut score of 785, we get 160. 160 divided by 5 equals 32. Therefore 80% of the way up from 785 to 945 is 913, and 913, then, can probably count as C1. To put the figure of 913 into context, it is interesting to note that the average of the ETS standard-setting in 2005, 880, and 945 is 915--almost exactly the same as the 913 arrived at by applying the 80% rule to 945.

This post hoc technique is also applied to the ETS cut-score for B2, 795, by subtracting the cut-score for B1, 550, and dividing the result, 235, into five increments of 47. The point at which a person has all of the B1 descriptors and probably 80% of the B2 descriptors is 740, which is just about exactly what the English program has been using as a minimum achievement benchmark for 2nd year English majors: 730. If one uses 77%, instead of 80%, in mapping, then 730 is the effective B2 cut score.

It is interesting to note that the reason the English program has been using 730 TOEIC as a benchmark is because that is the figure arrived at by *Mombukagakusho* as a cut for upper-intermediate level (**personal communication, Okada sensei, Osaka university**). This could be interpreted as independent evidence corroborating the need to apply the 80% rule to the ETS standard-setting, 2008.

2.2.4 ETS claim TOEIC cannot prove C2, yet top TOEIC scorers C2 by other indicators

The panelists of the 2008 ETS standard-setting study agreed that the TOEIC could not provide conclusive evidence of C2 for any skill. Looking at examples of the TOEIC and the set of descriptors for CEFR C2 reading, this seems a reasonable conclusion, in theory, in which case post hoc application of the 80% rule would seem impossible. However, in practice, is it reasonable to assume that everyone with real C2 level will always achieve a perfect TOEIC score? Would all native

university students always produce a perfect score? It seems unlikely, as a result of factors other than linguistic competence: People make mistakes due to fatigue, lapses in concentration, unfamiliarity with a test/test format, or unfamiliarity with a test subject bias, such as business English. Therefore, logically, even by ETS's standard, while a perfect TOEIC score presents no clear evidence for C2 level, a perfect or even near perfect score provides little or no evidence against C2 level. Practically speaking, it seems highly likely that even a slightly less-than-perfect TOEIC score may result when a person of real C2 level takes the test.

Moreover, according to the results of BULATS, Self-assessment (1 of 2) and Teacher assessment, it seems extremely likely that at least two of the end-of-term 1st year students had C2 level. Finally, as additional evidence for C2, it should be noted that each of the students in question has extensive experience in English-speaking environments.

Therefore, the second highest score, 965, is posited in Smith-mapping as entry to C2, which is 5.8 out of 6 using Smith's numbering: A1=1, A2=2, B1=3, B2=4, C1=5, C2=6. The person with 965 is considered to have all of the C1 abilities, plus about 80% of the C2 abilities. Notice, however, that nothing definitive can be said about the person with the highest score, 975, using this method, except that s/he is a little over the hypothetical 5.8 or 80% of C2. Because we do not know the TOEIC score that proves C2, or even if it exists, we cannot say what scores over 965 mean, except that they are closer to full C2, and may strongly suggest C2 but do not prove it.

Smith-mapping attempts to account for the BULATS, Self-assessment, and teacher assessment data, without violating the fundamental claims of the ETS's 2008 standard-setting study, which it uses as a base. It produces reasonably consistent results for this data, and may be a good heuristic for the time being. However, the sample is too low at each small cut to be statistically robust. Note also that there is nothing above C2 in the CEFR, so there can be no 6.2, 6.4, etc. which occurs in self-

assessment below full C2 when peoples have different levels in different skill areas or a varied "profile". A person who self-assesses as C2 in Reading and Listening, for example, and C1 in Writing, Spoken Interaction, and Spoken Production would get a 5.4 using the author's self-assessment system.

2.3 Taking all heretofore discussed into account: Smith-mapped table explained by column

Below, I explain what each column means in the following table, Table 4:

I. Here, the ETS claim that TOEIC does not provide proof of C2 is accepted. However, in this column 965 is posited as 5.8 or 80% of C2, and it is suggested that that may be good enough to count as C2. The cut score to prove C1 100% is taken from ETS's most recent guidance and is 945; however, 80% of that, 913, is suggested to be good enough. The cut score to prove 100% of B2 abilities is 795, and 80% is 740 (77% is 730).

II. Students' reported TOEIC scores (The average is 790; ETS's cut score for B2 is 785)

III. TOEIC score of students/subjects smith-mapped to CEFR using the author's numbering: A1=1, A2=2, B1=3, B2=4, C1=5, C2=6; each level is one higher than in the ALTE numbering)

IV. Self assessment of students/subjects (Average is 4.05; B2 is 4 in the author's numbering and indicates 80% or more of can-do's from Checklists are affirmed); B2 can-do's are mostly completed: some tough ones remain: See JALT paper results).

V. BULATS scores, plus teacher comments about whether it is likely higher or lower than the students actual competence. It is assumed that several students under-performed for one or more of several reasons discussed in section 3.4

VI. This shows the CEFR level of students/subjects according to the easier version of ETS's C1 cut score explained in recent guidance, together with the author's evaluation of that ETS-based CEFR level.

VII. This shows the CEFR level of students/subjects according to the harder version of the ETS cut scores that resulted from the standard-setting in 2008, together with the author's evaluation of the ETS based CEFR level,

Smith's numbering: A1=1, A2=2, B1=3, B2=4, C1=5, C2=6

See Table 4, Smith Mapping, below.

3. Possible test inaccuracies

3.1 TOEIC and BULATS: Listening and Reading only

The TOEIC and BULATS versions used above contained only listening and reading sections while the Self-Assessment contains five sections: Listening, Reading, Spoken Production, Spoken Interaction and Writing. For each subject, the scores from each of these five sections were averaged. This extra information in the Self-assessment scores above, not represented in TOEIC and BULATS, likely interferes with mapping from Self-Assessment averages to those tests and clear correlations were not found. Generalizations, about CEFR level based on TOEIC and BULATS scores are valid for all five CEFR skills only when a student's speaking and writing abilities are equivalent to the listening and reading, which is often not the case. In a future study, Separate Listening and Reading scores for TOEIC, BULATS and Self-Assessment should be collected and compared. BULATS Listening and Reading are compared to Self-Assessment L & R in Section 4 and adjusted based on teacher analysis.

3.2 TOEIC high:

1. Students studied for the TOEIC to varying degrees in a 1st year class
2. TOEIC minimum requirements in place: 680 1st year and 730 2nd
3. Used for employment
4. No British English bias

Table 4 Smith Mapping

I	II	III	IV	V	VI	VII
C2: 5.8 (965=5.8) 100% proof not on test	975 965	6 6	6/C2 4.6 (underestimated)	6/C2 (Yes) 6 (Yes brd) R: 86 L:92	5/B2 (wrong) 5 (wrong)	5/B2 (wrng) 4 (wrong)
C1: 5.6 (960)					5 (wrong)	4 (wrng)
C1: 5.4 (955)					5 (wrong)	4 (wrng)
C1: 5.2 (950)	950	5	6 (over estimated? Maybe; maybe not)	4 (wrong: oddly poor performance. American high school)	5 (maybe)	4 (wrng)
C1: 5 (945)	945	5	4.6 (undr estimated?)	5	5 (maybe)	4 (wrng)
C1: 4.8 (913-944) 80% or more of C1, so can be thought of as being within C1, range (80-99%)	930 915 905 900 890 890 885 885	5 5 4 (brdr) 4 4 4 4 4	3.2 (under estimated) 4.2 (under est.) 5 (border: o.k.) 3.8 4.2 3.8 4.8 (border: o.k.) 4.8 (border: o.k.)	4 (wrong) / border: o.k. 3 (wrong: poor perf.) 5 (maybe) 4 4 4 4 4	4 4 4 4 4 4 4 4	4 4 4 4 4 4 4 4
B2: 4.6 (881)	875 865 850	4 4 4	5.8 (over estimated) 3.2 (under est.) 4.6	4 3 (poor performance) 4	4 4 4	4 4 4
B2: 4.4 (849)	830 825 825 820	4 4 4 4	3.2 (under est.) 5.4 (over est.) 3.6 (under est.) 3.4 (under est.)	3 (poor performance) 4 5 (stale TOEIC?) 4	4 4 4 4	4 4 4 4
B2: 4.2 (817)	815 810 805 805 790	4 4 4 4 4	3.8 4.6 3.8 4 3.2 (under est.)	5 (stale TOEIC?) 4 4 4 4	4 4 4 4 4	4 4 4 4 4
B2: 4 (785)	785	4	4.2	3 (poor performance)	4	4
738-784 = 80-99% of B2, so these scores probably amount to B2	780 780 763 760 755 750 750 745 740 740	4 4 4 4 4 4 4 4 4 4	3 (under) 4.2 4.2 4.2 3.2 4 4.4 4.4 3.6 (border) 3.2 (border)	4 4 4 4 3 (crammed TOEIC?) 4 4 4 3 (border: o.k.) 4	3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong)	3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong) 3 (wrong)
B1: 3.79 (737) If we accept 77% of B2 as qualifying for B2 then 730 is good enough	735 730 715	3 or 4 3 or 4 3	4.6 (border) 4.8 (over est.) 2.6 (under est.)	4 (border: o.k.) 4 (border: o.k.) 3	3 (wrong?) 3 (wrong?) 3	3 (wrong?) 3 (wrong?) 3
B1: 3.6 (691)	680 655 650	3 3 3	3.6 4.2 (stale T) 4.8 (over est.)	3 4 (stale T) 3	3 3 3	3 3 3
B1: 3.4 (644)	610 600	3 3	4 (over est.) 3.4	2 (poor p.) 2 (poor p.)	3 3	3 3
B1: 3.2 (597)						
B1: 3.0 (550)						

3.3 TOEIC low:

1. Score can be "stale"/not current at time of survey.
2. Business English focus may be unfamiliar

3.4 BULATS low:

1. No preparation; completely unfamiliar format
2. No BULATS requirements in place (test not a "gatekeeper" in program)
3. Value for employment not yet clear to students
4. British English bias
5. Business English may be unfamiliar

3.5 BULATS high:

1. Exposure to British English: Top scorers have
2. Exposure to business English

4. Minimum Achievement Levels: set near finishing level of lowest students, by year

How should minimum achievement levels be decided? It depends on the norms of the institution and the country. At the University of California, in the U.S., about one third of 1st year students do not continue to 2nd year; in France the numbers cut at public universities are even higher. In Japan, however, the lower-level students are not let go, so the minimum achievement level is, in effect, set by them. The average level, however, is much higher, and the highest-level students are much higher still.

Table 5 Average level by source

Source of average	Average	B2 cut-score
TOEIC	790 (some "stale")	785
Self-Assessment using Swiss Version of Checklists	4.09 (tendency to under ass.)	4
BULATS	3.74 (tend. low: see above)	4 (According to BULATS, converted to author's numbering)
Teacher Assessment	4.3 (probably right given tendency to underestimate and some bad performances on BULATS	3.8 (according to Swiss Version of Self-Assessment Checklist instructions)

Although the average level at the end of 1st year is definitely B2, because a few students are still in B1 range at end of year one, B1 must be the minimum achievement level for 1st year. The average starting level of 1st year, according to self-assessment in April 2009 is B1; however, some students assessed as A2 in speaking and listening. Usually, these students only reach B1 by end of 1st year. With some fine-tuning of the curriculum (Majima & Smith 2008, Smith & Smith 2009) and/or the entrance exam to better filter out low-level students, it may be possible to raise the minimum achievement target. On the other hand, this year's 1st year students may all reach B2 level in terms of TOEIC in January/February, 2010. If so, and if future classes do the same, the minimum achievement level could be raised.

As for 2nd year English majors, the results of Self-Assessment will be coming in January and February. The author predicts that the average should be significantly higher in the B2 range, maybe even nearing C1. However, students with TOEIC scores already in excess of the benchmarks are unlikely to take TOEIC

again, and they are not scheduled to take BULATS. Nevertheless, the average of self-assessment has been remarkably accurate in terms of other measures, including TOEIC, BULATS and teacher assessment, so, baring the introduction of new variables, whatever the average of their reported scores is should be right. Nevertheless, if the average does climb to B2+ or even C1, there will still be some students just meeting or exceeding the 2nd year benchmark of 730, which arguably amounts to B2, so the 2nd year minimum achievement can be said to be B2.

5. Teaching Targets: set by year, above average starting level

The match between ETS's cut-score, 785, students' average TOEIC, 790, Self-assessment, 4.05, and BULATS, 3.74 (real level should be near 4, if causes for low scores removed) is extremely good, so there can be little doubt the average level of students at end of 1st year is B2. Therefore, the teaching target for 2nd year must be C1, and texts used should target C1. However, some of the least affirmed can-do's from B2 must still be treated in 2nd year. These have been identified by Smith & Smith (2009). Smith's WebCT version of the Checklists provides a valuable means of understanding what students believe they must learn in terms of

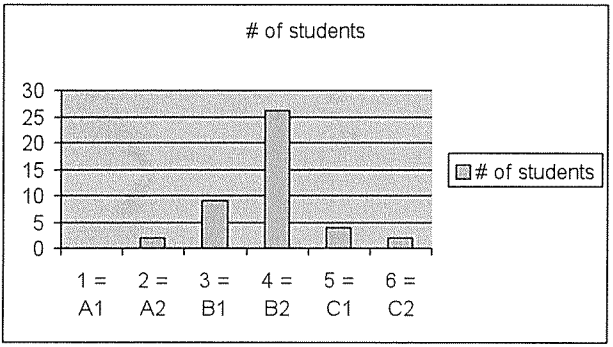

Figure 1 Number of students by BULATS-CEFR level

Checklist can-do's.

Teaching Targets must treat the needs of the majority of students—that is average level—as long as students are not divided into classes by level. As discussed above, most of the students were put at a disadvantage by the British English bias, complete lack of familiarity with the test etc., so it may be reasonable to assume that at least those students with an average score of 74 on BULATS (still B2), one point short of C1, are in fact just qualified C1. Similarly, the real level of a few A2 and B1 students on the borders is probably one level higher. Thus, the vast majority of students should be ranked B2 in terms of BULATS and about 10% should be within the C1 range, or higher. Therefore, the teaching target for 2nd year should be essentially C1 and its can-do's, plus the least affirmed can-do's from B2 (See Smith & Smith 2010).

With separate levels for English majors, language curriculum could better focus on the language needs specific to each, including those entering with B2 or even C1 level, perhaps allowing them to reach C2 by end of 2nd or 3rd year. To be properly challenged in 1st year, such students need materials and curriculum that target B2+, C1 or even C2. However, levels for majors is counter to tradition in the English program and may have negative side effects. As it is, teachers try to adjust the current curriculum for such students on a case-by case basis.

5.1 Suggestions given long B2 range in TOEIC and BULATS:

There is a long stretch of B2 in terms of TOEIC scores, 785-980+ according to ETS, June 2008, or 740 up to 913, according to Smith-mapping of ETS. Perhaps there is a need for a B2+ or C1- cut-score to give students an intermediate learning goal and teachers an intermediate teaching goal. Although the range in BULATS is not so large, a B2+/C1- cut might help motivate students along the way to C1. As it is, more than a few students may spend an uncomfortably long time getting through B2 and become discouraged. If C1 is roughly equivalent to EIKEN 1st

grade, then B2+ could serve a role similar to that of Pre-First Grade[13]. It will not be known until data is collected at the end of January if 2nd year students self-assess on average at C1 at the end of the term, or slightly below. If it is not C1, then B2+ or C1- should be established as a teaching target to distinguish the 2nd year teaching/ learning target from the 1st year target.

6. Under-estimators and over-estimators

Compared to TOEIC, teacher assessment and BULATS, the young men in this study tended to accurately assess or overestimate their abilities while the young women tended to accurately assess or underestimate them. It is my hypothesis (idea 1st introduced by Smith & Smith 2005) that this is due to a cultural tendency and that similar results will occur in future studies. Please see the table below.

Table 6 BULATS and Self-Assessment by Gender

(44 students supplying data in all fields)[14]

Reading	Reading	Listening	Listening
MO = 11/17	FO = 8/26	MO = 8/17	FO = 8/26
MU = 1/17	FU = 6/26	MU = 3/17	FU = 8/26
MM = 5/17	FM = 13/26	MM = 6/17	FM = 11/26

Table 7 BULATS vs. Self Assessment by Gender adjusted with Teacher Assessment

Reading	Reading	Listening	Listening
MO = 11/17 (7/17)	FO = (0/26)	MO = (6/17)	FO = (2/26)
MU = 1/17 (2/17)	FU = (6/26)	MU = (3/17)	FU = (8/25)
MM = 5/17 (8/17)	FM = (20/26)	MM = (8/17)	FM = (16/26)

6.1 Analysis

6.1.1 Listening: British bias etc. may cause border scores to be one level two low.

Of the female subjects, one is not Japanese, so for this Japanese cultural analysis, she is not included. In terms of listening, there is one case in which both BULATS and self-assessment are wrong, with the real score being most likely in between. In all the other cases, the so-called female over-estimators were right, and they under-performed on BULATS for any one or more of a variety of reasons discussed below. If so, only two women over-estimated, and we can conclude that by-and-large women estimate accurately or underestimate, in terms of listening. Taking this into account, for women, the Checklists for listening have some predictive power.

Males on the other hand, over-estimated in listening more than half the time, and rarely under-estimated. Future data will reveal the degree to which this trend continues. In the meantime, curriculum planners looking at male and female data can take this initial result into account when setting levels.

6.1.2 Reading Over-estimation: BULATS C1 & C2 requisite speed and vocabulary not on checklists

In terms of reading, the number of apparent female over-estimators grows little; however, in this case the over-estimation tends to be by two points. Are the women grossly over-estimating? Has their cultural tendency toward "*kenkyo*" changed here? No, probably not. Many of them can read high-level texts intended for native speakers, in some cases "with ease" and in many cases with the help of a dictionary, as suggested in the C1 levels of the CEFR Checklists. However, some of the C1 Checklist can-do's seem to indicate that one can take one's time and use a dictionary (see below), but test-takers cannot do these things when taking BULATS:

I can understand long complex instructions, for example for the use of a new piece of equipment, even if these are not related to my job or field of interest, provided I have enough time to reread them. I can read any correspondence with occasional use of a dictionary.

Moreover, the C2 can-do's in the reading Checklist do not include anything about speed. Given these facts, it is easy to understand how the female over-estimators self-assessed via the Checklists as C1 or even C2. However, it is highly probable that many who made such high self-assessments could not read the texts in the BULATS exam quickly enough in general and well enough without a dictionary.

This is especially true because the test focuses on business English, which is likely to be unfamiliar to students under twenty, and reportedly tends toward British English, which is also unfamiliar to most of the students, as British English can differ from North American English in terms of spelling, punctuation, tense, aspect and modality. If these differences were indeed present, they could well have caused many of the apparent over-estimators to under-perform.

Two of the B2 descriptors (see Appendix B) mention speed.

I can rapidly grasp the content and the significance of news, articles and reports on topics connected with my interests or my job, and decide if a closer reading is worthwhile.

I can quickly look through a manual (for example for a computer program) and find and understand the relevant explanations and help for a specific problem.

However, being able to read speedily is not mentioned in the C1 and C2 descriptors; in fact, rather the opposite feeling is conveyed, as above. This may be a reason why many students who felt they qualified as C1 or C2 in reading did not, according to BULATS. Although repeating a point in multiple levels may have its own

disadvantages in terms of making the Checklists, it may be advisable in future versions of the Checklists to let students know they have to read and understand quickly to reach C1 or C2, on a standardized test like BULATS.

Along the same lines, to improve the convergence of Self-Assessment ranking on the CEFR scales with BULATS ranking, it is advisable that BULATS disclose to stakeholders precisely what constructs, or range of constructs, from CEFR are being tested, how well those match the ELP checklists, and, if necessary what constructs/ can-do's in the Checklists and/or BULATS should be revised in order that a course of study guided by the Checklists be maximally efficient in leading students/a program toward ever better BULATS scores, and, for that matter IELTS scores.

7. Merits and Demerits of BULATS

7.1 Probable Merits of BULATS

BULATS gives a CEFR score that seems by and large correct, according to this study. There are some students who likely under-performed on BULATS; however, if an institution officially adopts BULATS, many of the reasons for underperformance should disappear. That is, students or employees can become familiar with the test and come to try their best on it. The British English bias is an advantage for only a minority of students in Japan. There are just a few cases where students may have over-performed on BULATS, in comparison to other data, but these can usually be explained away by a stale TOEIC score and/or underestimation of CEFR level, almost always by women, in self-assessment.

7.1.1 BULATS Merits compared with TOEIC:

7.1.1.1. Price

The price is comparable to TOEIC and low enough to take more than once.

7.1.1.2. Business/practical English

Both TOEIC and BULATS focus on business/practical English, which is useful for employers and students, considering that most Japanese employers tend to hire students without graduate degrees, preferring to conduct OTJ (On the Job Training).

7.1.1.3. CEFR score

Various attempts have been made to link TOEIC to the CEFR, but published cut-scores have fluctuated dramatically. The author may have developed a principled solution to TOEIC's recent and probably overly high cut scores, but corroborating evidence remains to be gathered. Despite poor performances by some students, the correspondences between BULATS scores and CEFR seem to be sound with no need for adjustment. This is not surprising considering that BULATS is approved by Cambridge ESOL.

7.1.1.4. CEFR range

While TOEIC may be a reasonably good instrument for distinguishing between B1 and B2 (especially if the cut is adjusted downward as explained above), it's makers, ETS, have yet to produce robust, unchanging, cut scores for C1 and any cut scores for C2. According to its 2006 standard-setting results, none of the 1st year students are C1 by the end of the 1st year, even those who studied abroad for three years in high school. According to the modified guidance, the three top-scoring students are C1, which is still too few according to the teacher's assessment. Therefore, if a business is interested in properly discriminating between candidates/ employees with B2 (upper intermediate), C1 (advanced/operational proficiency) and C2 (mastery), then BULATS is a better choice. The full CEFR range also makes it a good choice for use by any institution that recognizes the value of CEFR, the number of which is increasing day by day. Moreover, firms based in Europe are likely more familiar with BULATS than TOEIC. The full range also make it

potentially very useful for a university English program that refers to CEFR, such as that in the School of Foreign Studies at Osaka University. BULATS can be a little bit conservative in granting B1, B2 and C1 near the borders, according to Smith-mapping, and/or some students under-perform for the reasons discussed above. However, overall, BULATS can discern CEFR levels pretty well including the advanced levels. ETS, on the other hand, ranks virtually any TOEIC score as being in the B2 range or lower.

7.2 Possible Demerits of BULATS

First, students perceive a British English bias; maybe this should be adjusted for Japan and Japanese who tend to be much more familiar with North American English. The two highest scorers, C2, had extensive exposure to British English. One returnee from America performed lower than teacher, student and even ETS would predict.

Second, students taking BULATS may still need to take TOEIC if employers want to see a TOEIC score.

7.2.3. BULATS CEFR scale judgments that 2008 ETS (in some cases), the Cambridge Catalogue and the teacher (the author) would consider too low

Table 8

CEF rating by BULATS	TOEIC Score (ETS TOEIC cuts) C1 = 945, B2 = 785	Cambridge Catalogue TOEIC cut heuristic: C1 = 800, B2 = 700	Teacher's estimate of student's level in terms of CEF	Real rating likely to be at least one CEFR level higher than BULATS' rank
B1	785 TOEIC = B2	785 TOEIC = B2	B2	Yes.
B1	830 TOEIC = B2	830 TOEIC = C1	B2 high	Yes
B1	865 TOEIC = B2	865 TOEIC = C1	B2 high	Yes
B1	915 TOEIC	915 TOEIC = C1	C1 (within 80%)	Yes
B2	950 TOEIC= C1	950 TOEIC = C2	C1/C2 (high school in U.S. completely fluent)	Yes
B2	930 TOEIC= B2	930 TOEIC = C1	C1 (within 80%)	Yes

7.2.4. BULATS not typically used as an English gatekeeper at universities.

IELTS would solve this problem and deliver all of the advantages of BULATS except price, which make it impractical for frequent use. However, for a single use before study abroad, IELTS could be ideal as its cost is similar to TOEFL *iBT*. Moreover, representatives from Cal State Fullerton[15] recently commented that more and more students are taking IELTS because they feel it is easier to meet entrance requirements that way than TOEFL (Tomoko Y. Smith personal communication). I have yet to see additional data to corroborate that claim, but for students studying in a system that refers to CEFR, I would guess that it is true.

8. Final future considerations

Seeing the oddly high TOEIC scores not long after they were published, the author's instinct was to question these claims as it has been his experience that students with TOEIC scores well over 900 are likely to be C1, and students with exceptionally high TOEIC scores—perhaps 950 and up, display native-like speech within certain topical limits, which may be indicative of C2 level and/or may indicate that they possess what Hulstejn refers to as "core language proficiency".

The core of language proficiency restricts this knowledge and skill to frequent lexical items and frequent grammatical constructions, that is, to lexical items and syntactic constructions that may occur in any communicative situation, common to all adult NSs regardless of age, educational level, or literacy.

In first year, the two highest scoring students have extensive English background. In the highest scorer's case (TOEIC 975, self-assessment C2, BULATS C2) while growing up, the mother, Chinese, with university education in Hong Kong, and father, a Japanese professor of Australian History/culture, communicated with each other in English. However, around the time the student was a teenager, the mother switched to Japanese. Also, the student often visited Australia with the father and America with relatives of the mother. The 2nd highest scoring student (TOEIC 965, self-assessment C1/C2, BULATS C2) attended an international high school in the Netherlands where English was the mode of instruction. BULATS seemed to detect the native-like ability of these students when it selected these top two TOEIC scorers as C2. It should also be noted, however, that the top two scorers were exposed to British English, or varieties close to it.

On the other hand, one student who spent three years at high school in Canada (TOEIC 945, self-assessment C1/C2) who displays native-like speaking ability was

ranked by BULATS as C2 reading (84/90) and C1 listening (83/90). Perhaps she could have reached a C12 total if the listening section were in Canadian English. Moreover, BULATS ranked one student with excellent English ability (TOEIC 950, Self-assessment C2) and three years of high school in the U.S.A much lower than expected, BULATS C1 listening and B2 reading. That student also displays native-like speaking ability on limited topics in terms of pronunciation and intonation of American English; and, had the listening section been strictly North American English she would likely have been ranked C2. In terms of reading however, the BULATS rank of B2 indicates either an unusually bad performance, or non-native-like reading speed, vocabulary etc. These results may support the notion of "core language ability", which deals only with listening and speaking and not educational background/high level vocabulary etc. Future studies comparing separate listening and reading scores for Self-assessment, BULATS and TOEIC should help shed light on this matter. In the current study, separate TOEIC scores for listening and reading were not available.

In addition to the above questions regarding reading, a great many other students also performed much lower than expected in reading, and the probable causes of this may be the British English bias or BULATS testing some abilities the students lack. If it is the latter, maybe the can-do's in the Checklists need to be adjusted to better match BULATS; for example, "I can do the reading tasks described in this section quickly, and without a dictionary even when it involves British English and business English." This should produce better predictive power for the Self-assessment checklists' advanced reading section.

Regarding C1 and C2 speaking and writing, it is likely that many average native-speaker students with average educational backgrounds cannot affirm all of the can-do's in the Checklists. The high-level rhetorical skills they describe are, however, emphasized in a "good" traditional European or colonial education, most probably due to the influence of classical Roman and Greek culture on Europe.

Similarly, it is also likely that many Japanese high school graduates, and even university students, lack not only some C1 and C2 speaking and writing abilities in their native tongue, but also some B2 abilities because the Japanese educational system, at least according to many students and teachers I have spoken with, does not emphasize rhetoric: written/oral argumentation, formal debate and other speech communication skills. Japanese university English programs aiming to give students C1 level in all skills should take this into account, for it is unlikely that students can reach C1 level in terms of all the writing and speaking descriptors, without serious study of academic writing, speech and debate.

Eventually, if this study's results are corroborated in future studies, BULATS could contribute to the creation of a CEFR-based level system for general English education at universities and/or other institutions in Japan. To facilitate this, however, the British bias should be reduced and the relevance of BULATS in securing employment should be increased.

Finally, now that various means of assessment have been examined, including TOEIC, Self-assessment and BULATS, and reasonably consistent results found for this body of students, English majors in the School of Foreign Studies at Osaka University, future research concerning their English education should aim to find ever more effective/efficient teaching methods and materials for each year given the levels and abilities they need and value. Some recent suggestions for effective, CEFR-based, action oriented teaching, can be found in North, Babylonia February 2008 and Hans-Peter Hodel in the same issue. Also, further research as to what can-do's should be added or subtracted from the Swiss version of the ELP checklists for students with majors other than English should be examined (see Smith, Tomoko forthcoming)

Appendix A

Descriptors used in ETS 2006 Standard Setting

Speaking skills of just-qualified B2

Can give clear, detailed descriptions and prepared presentations attuned to the listener.

Can develop clear arguments with relevant support and examples on wide range of topics related to

fields of interest.

Can sustain conversation with degree of fluency and spontaneity.

Takes listener and cultural context into account.

Monologue causes no undue stress to listener.

Speaking skills of just-qualified C1

No strain on listener.

Expresses self fluently and spontaneously, almost effortlessly.

Uses idiomatic speech.

Uses precise and accurate grammar.

Can vary intonation and place stress correctly.

Can describe or present complex subjects (appropriately structured).

Shows flexible/effective use of language (humor).

Speaking skills of just-qualified C2

Effective and flexible communication with audience.

Can easily follow and contribute to complex discussion with all speakers.

Can express fine shades of meaning.

Can discuss abstract topics beyond own field.

Uses multiple registers appropriately.

Clear, well-constructed, smoothly flowing arguments.

Demonstrates full confidence in speaking.

Writing skills of just-qualified B2

Can produce a clear, detailed text essay.

Can argue for/against a position.

Can describe advantages/disadvantages.

Variety of subjects related to field of interest.

Easy to follow the structure but cohesion may be lost at times.

Texts are based on standard patterns.

Writing achieves clear, effective communication.

Can synthesize.

Uses informal/formal register.

Writing includes vocabulary related to field and good terminology.

Can write compound and complex sentences that will not lead to misunderstanding and do not impede

meaning.

Adapts standard format to personal needs.

Writing skills of just-qualified C1

Produces longer, well-structured and well-developed texts.

Uses language flexibly; mostly accurate

Elaborates to some degree.

Writes on complex subjects, with some degree of effort (time, dictionary, aids)

Can distinguish between formal and informal.

Uses efficient style (less wordy).

Writing skills of just-qualified C2

Produces clear, smoothly flowing, complex texts in an appropriate and effective style.

Writing is more natural/spontaneous.

Can write about all subjects.

Writing includes finer shades of meaning and frequently includes idiomatic expressions.

Produces smoothly flowing sentences/paragraphs; complex, extended texts.

Writing is characterized by range-appropriate style/register.

Uses cultural reference (e.g., politeness).

Takes reader's needs into account.

Can write complex, extended text.

Maintains consistent, highly grammatical control of complex language.

Makes few errors, if any.

Panel 1 and 2 Indicator Summaries of Language Skills Defined by the CEFR

Listening skills of just-qualified B2

Can understand standard speech on most topics.

Can use macro-structural clues to check for overall understanding.

Can grasp the main points of academic lectures.

Can understand radio and television.

Can understand speech from native speakers directed at him/her most of the time.

Can understand extended speech and complex arguments; requires explicit markers.

With some effort can catch most native-speaker discussion.

Can understand standard dialect delivered at normal speed.

Listening skills of just-qualified C1

Can understand extended speech on abstract unfamiliar topics (e.g., lectures).

Can understand enough but may need clarification.

Can follow most speakers.

Unfamiliar accents can cause difficulties in comprehension.

Does not require explicit markers.

Can recognize a wide range of idiomatic speech.

Can listen between the lines; can infer implied meaning.

Listening skills of just-qualified C2

Has no difficulty understanding any kind of standard spoken language, even when delivered at fast

native speed.

Will need time to adjust to nonstandard or colloquial speech.

Reading skills of just-qualified B2

Can read with a large degree of independence.

Can read texts in a wide range of professional topics (may need dictionary).

Has a broad, active vocabulary but has difficulty with low-frequency idioms.

Understands articles written from a stance (opinions and attitudes).

Can scan complex texts, locating relevant details.

Shows inferencing ability at macro level (text level.)

Reading skills of just-qualified C1

Needs to re-read; more effort required than C2 for complex, extended text in all fields of interest.

Identifies or infers opinion, intention, feelings of writer.

Reading skills of just-qualified C2

Reads practically all types of texts and styles, from most formal to highly colloquial.

Can critically interpret both explicit and implicit meaning.

Appendix B

Swiss Version of the ELP Self Assessment Checklists, B2-C1, in Reading and Listening

http://www.coe.int/T/DG4/Portfolio/documents/appendix2.pdf

B2 Reading

I can rapidly grasp the content and the significance of news, articles and reports on topics connected with my interests or
my job, and decide if a closer reading is worthwhile.

I can read and understand articles and reports on current problems in which the writers express specific attitudes and
points of view.

I can understand in detail texts within my field of interest or the area of my academic or professional specialty.

I can understand specialised articles outside my own field if I can occasionally check with a dictionary.

I can read reviews dealing with the content and criticism of cultural topics (films, theatre, books, concerts) and
summarise the main points.

I can read letters on topics within my areas of academic or professional specialty or interest and grasp the most important points.

I can quickly look through a manual (for example for a computer program) and find and understand the relevant
explanations and help for a specific problem.

I can understand in a narrative or play the motives for the characters' actions and their consequences for the
development of the plot.

C1 Reading

I can understand fairly long demanding texts and summarise them orally.

I can read complex reports, analyses and commentaries where opinions, viewpoints and connections are discussed.

I can extract information, ideas and opinions from highly specialised texts in my own field, for example research reports.

I can understand long complex instructions, for example for the use of a new piece of equipment, even if these are not

related to my job or field of interest, provided I have enough time to reread them.

I can read any correspondence with occasional use of a dictionary.

I can read contemporary literary texts with ease.

I can go beyond the concrete plot of a narrative and grasp implicit meanings, ideas and connections.

I can recognise the social, political or historical background of a literary work.

References

Council for Cultural Co-operation, Education Committee, Modern Languages Division, Strasbourg 2001: Common European Framework of Reference for Languages: Learning, teaching, assessment; Cambridge University Press

Heidi Byrnes, Associate Editor, 2007 "Perspectives" [on the CEFR]; *Modern Language Journal* 91 pp. 640-685

Hodel, Hans-Peter 2008. Intégrer le plan d'études dans l'enseignement / apprentissage; *Babylonia* 2 pp58-59

Hulstijn, Jan H. 2007. The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency; *Modern Language Journal* 91 pp. 663-667

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 and Antonio Smith 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

North, Brian. 2008. The Relevance of the CEFR to Teacher Training. *Babylonia* 2 pp55-57

Smith, Antonio 2009. "CEFR Self-Assessment Checklists' Impact on a Japanese National University's English Program"『英米研究』Volume 33 Osaka University March, 2009 pp125-138

Smith, Antonio and Tomoko Smith 2005.「大阪外国語大学英語専攻におけるCEFRを基盤にした 語学能力の到達度評価について」The first meeting of LPRG (Language Pedagogy Research Group) Osaka University of Foreign Studies

_____. 2009. "CEFR Self-Assessment Checklists' Impact on a Japanese National University" JALT 2009, 35th Annual International Conference. Grandship Shizuoka

Tannenbaum, Richard J. and E. Caroline Wylie 2008. *Linking English-Language Test Scores Onto the Common European Framework of Reference: An application of Standard-Setting Methodology.* Princeton, NJ: ETS

Acknowledgements:

This study was made possible thanks to funding provided by STEP and STEP's free administration of STEP BULATS to students in the School of Foreign Studies in April 2009. Much thanks to Okada sensei and Nakamura sensei of the English program in the School of Foreign Studies, Osaka University, for assisting in the administration and processing of STEP-BULATS. Much thanks to Okada sensei and Kishi sensei, English Program Chair for arranging a FD meeting in which to present this research. Much thanks also to Prof. Junko Majima of the Japanese program for securing funds[16] for research assistant Chiho Sakurai who in collaboration with Dr. Tomoko Smith constructed the Japanese-English bilingual WebCT version of the Self-Assessment Checklists according to my specifications. Much thanks to Chiho Sakurai and Dr. Tomoko Smith for their excellent contributions.

Notes

1) Common European Framework of Reference for Languages: Learning, teaching, assessment (2001)

2) Modern Language Journal 91 (2007)

3) See section 2.1.2

4) http://www.coe.int/T/DG4/Portfolio/documents/appendix2.pdf

5) http://www.ea.toeic.eu/uploads/tx_etsfreeresources/CEFRL_complete_study_02.pdf

6) ETS TOEFL Research Reports, R-R 80, November 2005: Mapping English Language

Proficiency Test Scores onto the Common European Framework, Richard J Tannenbaum and E. Caroline Wylie. http://www.ets.org/Media/Research/pdf/RR-05-18.pdf

7) ETS standard-setting 2006, published 2008

8) http://www.etscanada.ca/teachers/compare.php

9) http://www.etscanada.ca/teachers/compare.php

10) http://www.cambridgeesol.org/home/university-college.html

11) https://www.cl.etsglobal.org/fileadmin/free_resources/German%20website/Germany/Products/TOEIC/CEFR_Mapping_TOEIC_LR_TOEIC_SW_Bridge_2008.pdf

12) For another 2008 ETS chart with the same C1 cut see below http://www.ea.etsglobal.org/fileadmin/free_resources/ETS_Global_master/TOEIC_L_R_can-do_table.pdf

13) Thank you to Kishi sensei, English Program Chair, for the Eiken analogy; personal communication

14) MO: Male over-estimator, MU: Male under-estimator, MM: Male estimation matches BULATS FO: Female over-estimator, FU: Female under-estimator, FM: Female estimate matched BULATS.

15) Dr. Harry L. Norman, Dean and Ms. Lisa Xue, Director of International Programs University Extended Education, California State college, Fullerton

16)「言語教育における評価研究プロジェクト」の研究活動に必要な研究支援業務（Osaka University）