Accuracy in Speaking Self-Assessment among Japanese-speaking English Learners and Its Implications

Scott Smith

Abstract

This paper investigates the accuracy of self-assessment of speaking proficiency among undergraduates at a university in Japan. The value and utility of criterion-referenced self-assessment is discussed in light of dominant cultural discourses emphasising the importance of modesty, moderation and group harmony, which encourages Japanese learners of English to self-assess towards norms of "average" or "below average". The importance of criterion-referencing is discussed, and limitations in the data and criteria used in the study are explored. The self-assessment protocol is described, and the results of the self-assessments are compared with scores from assessments by trained and experienced raters conducting simulated IELTS speaking tests. The reliability of these expert assessments is explored using a measure of inter-rater reliability described in the paper. The study finds that the assessments were within the boundaries for inter-rater reliability in 10 of 11 cases. The data also suggests, in line with the literature, that self-assessment can be reliable – 8 out of 10 completed self-assessments were within the same boundaries – but becomes problematic at lower proficiency levels due to increasing inaccuracy. The implications of accurate self-assessment are discussed, consideration is given to how knowledge of one's own competences might interact with assessment and motivation, and the practical, affective and ethical ramifications of self-assessment and the challenges it presents to teachers, learners and curricula are examined. The paper concludes with the observation that further investigation of the reliability, scope and purpose of assessment is necessary.

Key terms: assessment, self-assessment, inter-rater reliability, criterion-referencing, IELTS, self-regulated learning

Literature Review

The ability to perform accurate self-assessment is a core component of self-regulated learning (Clark, 2012), which may be more familiar to English teachers as learner autonomy. Under models of self-regulated learning, self-assessment is one strand of feedback in a model of formative assessment along with feedback from both peers and teachers. Teachers are well placed to help learners who know and understand their own abilities, the goals of the learning encounter they are engaged in, and hold positive beliefs about their self-efficacy – the belief that they can take actions which will contribute to a positive outcome (Gallagher, 2012, p. 314). Previous studies, such as that by Muñoz and Albarez (2007) and the meta-analysis by Blanche and Merino (1989) indicate that self-assessment is positively linked to motivation, and that learners can self-assess with relative accuracy. However, learners both over- and underestimate their skills. Additionally, accuracy varies positively with proficiency, and the self-assessment tools used require careful design. Culture may play a significant confounding role in such studies with Japanese learners, as dominant discourses in Japanese culture emphasise modesty (謙遜 – *kenson*), moderation (控えめ – *hikaeme*) (Brown, 2004) and group harmony (和 – *wa*). These have been shown to encourage learners – regardless of their actual ability – to report their ability to learn a foreign language as "average" or "below average" (Brown, 2004). To appropriate the Japanese proverb, the nail that sticks up will hammer itself down.

Two additional concepts further complicate discussion of norm-referenced self-assessment in the Japanese context, and are related to *wa*: *Honne* (本音) – behaving in ways and saying things that represent one's actual beliefs, dispositions and attitudes; and *tatemae* (建前) – behaving in ways and saying things that one believes to be appropriate in the particular social circumstance. These are particularly troubling concepts in the investigation of self-assessment, since the decision to say what a person believes is expected (rather than proffer their real opinion) is a private one, and seems to be linked to avoiding disapprobation rather than any misalignment between self-perception and reality (Yamagishi *et al.*, 2012). Therefore, establishing the veracity of participants' reports in such studies is dependent upon an unverifiable factor. Unfortunately, a lengthier discussion of cultural variation in the

epistemology of self-assessment is not possible in this paper. Suffice it to say that it is an undercurrent in explorations of self-assessment in the Japanese context.

It is important to note, however, that Brown's paper sets out to elucidate the roles of Japanese learners' beliefs about their English learning abilities through self-assessment, and that the investigation is therefore framed by its reliance on a tool that asks learners to assess their abilities relative to others, rather than by reflecting on descriptors of linguistic performance such as the CEFR 'can do' statements (Council of Europe, 2000), or the assessment criteria for the IELTS (IELTS, 2014a; 2014b; 2014c) and TOEFL iBT speaking tests (ETS, 2004). Clark (2012, p. 227) argues that "teachers should, as the starting point, de-emphasize social comparison and depersonalize feedback," as comparing participants in a learning group and the potential "construal of low attainments as indicants of inherent personal deficiencies erodes a sense of efficacy" (Clark, 2012, citing Bandura, 1997). He also states that "sharing learning intentions and identifying clear assessment criteria is the *sine qua non* of formative assessment." (Clark, 2012, p. 210) Thus, the limitations of comparative self-assessment and its potentially negative impacts on learners' beliefs about their self-efficacy form a compelling case for the use of criterion-referencing in assessment.

The use of criterion-referencing for self-assessment offers a means to address the confounding effect of normative pressure and the damaging effects of comparison in assessment. Nevertheless, this is still a relatively understudied area in Japanese EFL. Searches of the literature on the topic have revealed a limited amount of research into criterion-referenced self-assessment by Japanese-speaking English learners, including two papers from Runnels (2013; 2014), focussing on validation of the CEFR-J (cefr-j.org, 2014), a derivative of the CEFR (Council of Europe, 2000) localised for Japanese learners. However, Runnels' 2013 paper does not involve focused assessment of individual learners. Instead, participating teachers were asked to decide whether they were confident that 80% or more of their students could perform the descriptor. This aspect of the study found no correlation between the teachers' assessments and those of the students. There are, then, questions as to whether teachers – whose primary interest in teaching a class is arguably in forming constructive developmental relationships with their learners – can be reliable assessors of their own charges (Bérešová, 2011), an issue discussed in more detail later. As a final consideration, the CEFR and the CEFR-J are general competence descriptors, intended "for the *elaboration* of language syllabuses and curriculum guidelines, the *design* of teaching and learning materials, and the assessment of foreign language proficiency." (Council of Europe, 2014, emphasis mine). As such, an unelaborated CEFR/CEFR-J is a general document without a communicative context. This may make it difficult to judge competence without directly assessing individuals in a contextualised communicative interlude.

Aside from the CEFR-J, Yoshizawa (2009) investigated 'can do' statements derived from ETS' *Can Do Guide* (ETS, 2013), which were intended for deployment with younger learners with little-to-no experience of the contexts in which the business-oriented language used in the TOEIC might be found. She reported that there was a correlation between the learners' self-assessments and the difficulty of the reading descriptors, although she found no correlation for listening.

The studies by Runnels and Yoshizawa are suggestive, but nevertheless limited in their applicability to the context in which this research was undertaken: They focus on general English, rather than the linguistic and cognitive competences involved in preparation for entry into academic programs delivered in English. In short, the constructs were not intended to apply to the specific institutional context.

This paper, then, is the result of, and a response to, the lack of research into the accuracy of criterion-based self-assessment of productive skills in the Japanese tertiary context. There are explanatory factors for this absence: Primarily, experience shows that assessing productive skills is both time-consuming and difficult. For example, the IELTS, a widely-used and well-recognised proficiency test of English for academic purposes, involves a spoken interview of 11-14 minutes (Cambridge IELTS 8, 2011, p. 6). Undertaking a writing test such as the IELTS academic writing module requires an hour of the participant's time, and the writing produced should total at least 400 words over two scripts. (Cambridge IELTS 8, 2011, p. 5) Thus, speaking and writing cannot be assessed quickly.

Additionally, assessment requires practical training which includes principled feedback on assessor performance. In the Japanese university context, this is problematic. Practical teaching qualifications which might allow a trainee to explore formal assessment practices through action research, such as in Module 3 of the Cambridge Delta (Cambridge English, 2014) or the Trinity Diploma in TESOL (Trinity College London, 2007), are often ignored in favour of publications and academic qualifications with little-to-no practical component (Lowe, 2012), and opportunities for assessor training in the country are limited. Furthermore, common introductory qualifications, for instance the CELTA, focus on the practical skills involved in teaching and informal assessment such as concept checking and answer checking (CELTA Syllabus and Assessment Guidelines, 2010). Additionally, as assessment of receptive skills is markedly simpler than for productive skills – to the extent where it can be (and often is) automated – productive skills have received rather less attention than receptive. For these reasons, it was decided to undertake a small-scale study comparing students' self-assessments of their speaking skills with the assessments of trained and experienced assessors.

As the task of researching and developing assessment resources is complex and time-consuming, it was concluded that it was appropriate to adopt and adapt an existing framework. Additionally, assuming a framework enabled the adoption of characteristic tasks, negating the need to independently develop a test for the assessment construct. In order to satisfy the requirement that the test reflect the academic orientation of the English program at the institution, it was considered whether the constructs assessed passed a simple test of face validity: Were they academic in orientation?

Such deliberations reduced the number of candidate tests to two – TOEFL iBT speaking and IELTS speaking. As I am more familiar with the IELTS than the TOEFL, both in terms of the format of the test and the construct assessed, I decided to employ a derivative of the IELTS speaking test. It is worth noting at this time that later investigations into this topic ought to employ assessments derived from both the IELTS and from the TOEFL iBT to better investigate criterion validity.

Research Questions

The theoretical and practical considerations led to the research question, which is based on a hypothesis derived from the review of the literature.

Research Question

How well do students' self-assessments of their speaking skills, performed using a framework derived from the IELTS speaking criteria, correlate with the assessments of expert assessors using the IELTS public speaking scale?

Hypothesis

The participants can self-assess accurately, although linguistic competence and unfamiliarity with both the descriptors and self-assessment in general will lead to discrepancies between self- and expert assessor ratings.

Method

The Assessment Instrument

The instrument was structured to match the IELTS speaking test as closely as possible. Therefore, it is prudent to give a brief overview of the format and content of the speaking test before proceeding. For a fuller description, refer to www.ielts.org/teachers.aspx. Appendix 1 gives an overview of the timings and contents of the speaking test, and Appendix 2 provides a brief account of the four domains of competence rated.

The IELTS speaking test is a divergent, criterion-referenced test of English, in that the award of a score is not based on correct or incorrect answers, but rather reached through fulfilling topic-agnostic competence descriptors in four domains of language: fluency and coherence (FC), lexical resource (LR), grammatical range and accuracy (GRA) and pronunciation (Pron). These are outlined in more detail above in Appendix 2. The band scores in each domain range from 0 to 9. Zero is only apportioned

to candidates who do not attend or otherwise fail to complete the test. A band score of 9 represents well-educated native speaker-like competence in the language. Indeed, a native speaker of English may well fail to achieve a band 9. At the other end of the scale, bands 1 and 2 are below the level which the majority of learners of English enrolling in university in Japan have achieved.

While discussion of the highest and lowest bands of the IELTS speaking module is interesting, it is beyond the scope of this paper, except to illustrate that the highest and lowest bands fall outside our area of concern, since very few learners (and no participants in the study) fulfil the criteria. For these reasons, the bands 0 to 2 and 8 and 9 were excluded from the self-assessment tool, which had the additional benefit of simplifying it for the participants. The student self-assessment scores were recorded as IELTS band-equivalent scores when input, although they do not appear as such on the self-assessment form. Indeed, no numerical values appear on those forms at all in order to mitigate against learners aiming for "below average" (Brown, 2004) by selecting the median score.

With regard to simplification, even a brief glance at the public speaking criteria (IELTS, 2014a) should illustrate that the language employed is technical. Therefore, simplification of the language, in order to make it more accessible to learners was necessary. The simplified criteria are laid out in Appendix 3. Despite simplification, it was still necessary to devote time to explaining key terms to the participants and ensuring that they understood them in order to attempt to address the injunction that "teachers should model forms of discourse which support the description of assessment criteria [which] meet the needs of the students" (Clark, 2012, pp. 7-8). Explanation and scaffolding took around an hour in plenary and peer-to-peer activities. Where possible, the structure of the sentences was rearranged in order to make criteria more performance-oriented. This involved paraphrasing and replacing technical language, rephrasing sentences with "I can" statements, and generally using the first person in order to make the descriptions more concrete. This does raise some issues regarding reliability, which will be discussed later in some detail along with other limitations.

Participants

All participants completed an informed consent form where they indicated that they agreed to their voice being recorded and for their data to be used in the writing of the paper. The form also included a section allowing the participants to explicitly state that they permitted sharing of their anonymised voice data within the institution for training and development purposes. All participants signed and returned both sections of the form.

The participants numbered 13 volunteers. Encouragement to participate took the form of prepaid cards and vouchers for major Japanese online retailers, valued at ￥1,000 and purchased from personal funds, although this was not offered as an incentive to self-assess accurately (Yamagishi *et al.*, 2012). At the time, the participants were enrolled in the institution's capstone English program for students who indicated that their Japanese was stronger than their English in their application to the university. Participants came from two countries – nine from Japan and four from China – although nationality was not used in subsequent analysis of the data. One volunteer did not complete the self-assessment, as she felt she did not understand the criteria. However, she was happy to continue to be represented in the study, and her perception that she didn't understand the criteria well enough is directly relevant to this paper. Therefore, her data was retained. Two other students later withdrew from the study due to difficulties participating in the speaking assessment, and thus their data was removed, leaving 11 participants – seven from Japan and four from China.

Procedure

Self-assessment

Participants received a copy of the self-assessment criteria (Appendix 3) a week before the expert assessment took place, and both the participants and the researcher explored the meanings of the descriptors. On the day of the self-assessment, 30 minutes was set

aside in order for the participants to read the criteria and their notes again, and identify their level. To further ensure clarity of comprehension, I was on hand during the self-assessment process to provide support and scaffolding in using the criteria. I neither provided the participants with clues as to where their level might be, nor indicated my perception of the class mean score.

Once the self-assessment had been completed, I collected and stored the sheets in such a way that I remained unaware of the self-assessments until all interviews had been conducted, and scores apportioned, reviewed and assessed for reliability.

Trained Assessment

The materials for the interviews were derived from the IELTS 8 book of past IELTS papers (Cambridge IELTS 8, 2011). For a list of the questions and tasks used, see Appendix 4. Three components from section 1 were chosen, and Parts 2 and 3 were drawn from the same question paper to ensure that they were thematically linked as described in Table 1. I administered the interviews to the timings of the test, and followed the same test format, inasmuch as that is possible in simulation. The interviews were recorded in order to both resemble the real test and to allow for me to listen to the interviews again at a later date. I assigned scores immediately after each test was completed using the IELTS Public Band Descriptors for speaking (IELTS, 2014a).

Objectivity of Teachers as Assessors: Ensuring Inter-rater Reliability

Major issues have been identified in teachers assessing learners for whom and to whom they are responsible, with Bérešová (2011, p. 186) concluding that "it is possible to state that feedback from teaching in class influences teachers' judgements." Runnels (2013, p. 4) goes further, suggesting that we are "incapable" of forming an accurate picture of the competences of our learners. In order, therefore, to minimise these issues, the assessments were performed again six months after the volunteers had left the researcher's class, using the recordings and without referring to the previous scoring. Once the scores had been determined, they were compared with the original grades, and the more recent grades were favoured where discrepancies existed. To further ensure reliable grading, the help of three other suitable assessors was enlisted. They were provided with between three and five recordings each and rated them independently using the same tool. The results were then collated and analysed. The spread in assessor scores was calculated, as was the mean score in both assessments, which was derived by calculating the mean of the researcher's and the independent assessor's score and rounding down to the nearest 0.5 where the sums produced 0.25 and 0.75 averages. This arithmetic was performed in order to produce overall scores that follow the IELTS format, wherein there are only full band (x.0) and half band (x.5) overall scores. While total agreement in scores would be ideal, Kuiken and Veder (2014, p. 281) report that, in an exploration of rater consistency, "raters differed significantly in their views on the importance of the various rating criteria," which strongly suggests that perfect agreement between human raters is difficult to achieve. Thus, I consider agreement within a half band for the overall score, and within one band for each sub-score to be an acceptable margin of error for the purposes of this paper, and also extend this principle to conclusions drawn from the comparisons of self-assessment and trained assessor scores.

Results

Inter-rater Reliability

The results for inter-rater reliability are shown in Table 1. With one exception, the overall scores awarded to each learner are consistently within 0.5 bands, suggesting inter-rater reliability. It is worth noting that the independent assessors scored lower than the researcher did on five occasions, while he scored more strictly on three. This perhaps reflects the difficulties of assessing one's own students. On one occasion – Student 8 – there is a spread of 1 full band. Both the researcher and independent assessor agree on the pronunciation score, although the independent assessor rates fluency and coherence, lexical resource and grammatical range and accuracy one band lower than the researcher. In light of the previous discussion regarding objectivity of teacher-raters, I suggest that the independent rater's score take precedence. With this one exception aside, both sets of assessors' scores are within the 0.5 spread discussed above, and there are no instances where the evaluations in a particular domain differ by more than one point.

Therefore, I conclude that the expert ratings in this paper are broadly reliable assessments of the student-participants' spoken English.

Student	Researcher's Scores						Independent Scores						Mean Score	Spread
	FC	LR	GRA	Pron	Score		FC	LR	GRA	Pron	Score			
1	5	6	5	5	5.0		5	5	5	5	5.0		5.0	0.0
2	6	6	6	6	6.0		6	6	6	6	6.0		6.0	0.0
3	6	6	6	6	6.0		6	5	6	6	5.5		5.5	-0.5
4	4	5	4	5	4.5		5	6	5	5	5.0		4.5	0.5
5	4	5	5	5	4.5		4	4	4	5	4.0		4.0	-0.5
6	3	3	4	4	3.5		3	3	3	3	3.0		3.0	-0.5
7	6	6	7	7	6.5		6	6	7	6	6.0		6.0	-0.5
8	5	5	5	5	5.0		4	4	4	5	4.0		4.5	-1.0
9	4	5	5	5	4.5		5	5	5	5	5.0		4.5	0.5
10	6	6	5	6	5.5		6	5	6	6	5.5		5.5	0.0
11	5	5	5	6	5.0		6	5	6	6	5.5		5.0	0.5

Table 1: The researcher's and independent assessors' scores

Reliability of Self-assessment

While the data sample is too small to meet criteria for statistical significance, there appears to be a correlation between the scores awarded by the assessors and self-assessed scores. With this caveat in mind, the hypothesis *seems* to be supported by the data collected. Eight out of eleven students' self-assessments fell within the 0.5 band boundary deemed acceptable for the purposes of this investigation. However, it is not possible to claim that this result has any predictive power. A further, larger-scale study involving a representative sample of learners in an English program is called for.

Correlation of Average Scores

The data can be viewed in Table 2. Overall, there was a slightly closer correlation between the mean of the assessors' scores and the self-assessments than there were between the individual assessors' scores. All correlations were within the previously discussed 0.5 band spread for eight of the eleven participants, with no spread for five of the eight. Those results include the student who did not self-assess (Student 9). If her data is excluded, the correlation rises to eight out of ten students' self-assessments falling within the 0.5 band overall spread.

Student number	Researcher Score	Independent Assessor Score	Mean Score	Self-Assessment Score	Spread from mean	Spread from researcher	Spread from independent
1	5.0	5.0	5.0	5.0	0.0	0.0	0.0
2	6.0	6.0	6.0	6.0	0.0	0.0	0.0
3	6.0	5.5	5.5	5.5	0.0	0.5	0.0
4	4.5	5.0	4.5	6.0	1.5	1.5	1.0
5	4.5	4.0	4.0	4.5	0.5	0.0	0.5
6	3.5	3.0	3.0	5.0	2.0	1.5	2.0
7	6.5	6.0	6.0	5.5	-0.5	-1.0	-0.5
8	5.0	4.0	4.5	4.5	0.0	-0.5	0.5
9	4.5	5.0	4.5	0.0	N/A	N/A	N/A
10	5.5	5.5	5.5	5.0	-0.5	-0.5	-0.5
11	5.0	5.5	5.0	4.5	-0.5	-0.5	-1.0

Table 2: Self-assessments and spread from the mean score of the first and second assessors' scores

Non-correlating Average Scores and Sub-scores

A table breaking down the self-assessment scores discussed here is presented in Table 3. Apart from Student 9, who did not complete a self-assessment, of the three students whose overall scores showed a significant discrepancy (greater than 0.5 from mean), the largest spread was for Student 6 – two bands from the mean score, or 1.5 bands from the researcher's scoring. This inconsistency is large – the difference between a beginner/elementary learner (CEFR A1/2) and a strong intermediate learner (CEFR B1), which raises important questions for self-assessment of proficiency among elementary learners of English. The remaining student whose self-assessment scores fell outside the 0.5 boundary (Student 4) produced a "spiky profile". That is to say that the scores she assigned herself in the four domains demonstrated a spread of two points between the highest and lowest sub-scores. For example, she selected the equivalent of an IELTS band 7 for fluency and coherence – entering the "advanced" level – but 5 for lexical resource, or intermediate. Student 4 was at a lower proficiency level among the students taking part in the study, with a mean score of 4.5. She, like Student 6, rated her abilities as significantly higher than the assessors did, with a divergence of three bands from the researcher's score and two bands from the independent assessor's score for FC.

Student Number	FC	LR	GRA	Pron	Score
1	5	5	6	4	5.0
2	6	5	6	7	6.0
3	6	5	5	6	5.5
4	7	5	6	6	6.0
5	5	5	3	6	4.5
6	5	5	5	5	5.0
7	6	5	6	6	5.5
8	4	4	4	6	4.5
9	0	0	0	0	0.0
10	5	5	6	4	5.0
11	5	5	4	5	4.5

Table 3: Self-assessment scores, in the four domains of fluency and coherence (FC), lexical resource (LR), grammatical range and accuracy (GRA) and pronunciation (Pron), with the overall score given on the right

Non-correlation of Sub-scores among Correlating Average Scores

Among the participants who self-rated to the standard of the assessors' average scores, there was no obvious pattern to the single band discrepancies in the four domains. Student 10's score was part of a spiky self-assessment profile. He rated himself as less proficient than the researcher did overall (5.0 against 5.5), with a discrepancy of two bands appearing in his evaluation of his pronunciation, although his overall score was within 0.5 of the independent assessor's and mine. This one example aside, no obvious pattern emerges in this data.

Discussion

Implications for Self-assessment among Lower Level Students

Both learners whose self-assessments diverged significantly from those of the expert assessors were rated as having achieved lower proficiency levels (mean scores 3.0 and 4.5). The participant in the study with the lowest level of English – Student 6 – provided the most erroneous assessment. Indeed, her self-assessment was flat: FC: 5, LR: 5, GRA: 5, Pron: 5, whereas the self-assessment of the other participant whose score fell outside the 0.5 band boundary showed spikes. I surmise that the inaccuracy in Student 6's self-assessment is at least partly due to her not understanding the meanings of the criteria, despite the support and scaffolding she received. This echoes Blanche and Merino's (1989) finding that lower-proficiency learners self-assess most inaccurately.

This tendency to inaccuracy in self-assessment among lower level learners raises important questions for educators wishing to implement self-assessment as part of a reflective, student-centred approach to teaching and learning. Most obviously, the issue of how to frame and structure the self-assessments to maximise accuracy (and therefore utility) becomes an issue. While developing a response to this issue and evaluating its efficacy is a worthy and interesting topic, it is deserving of its own research project. In the context of a short discussion in a short paper, I suggest that self-assessment for lower-level learners be based on achievement in a bank of representative tasks, in line with Blanche and Merino's finding (1989) that self-assessment is more accurate when connected to tasks learners have performed. It may also be the case that learners are more accurate when performing L2 self-assessment in their native languages. To ease comprehension of the criteria, it may therefore be worthwhile to translate them, a procedure not undertaken for this project. The above considerations also recall Runnels' point that "training students on self-assessment [is] likely [...] required if the CEFR-J is to be used for measuring language learning progress" (Runnels, 2013, p. 3).

General Implications of Self-assessment

This section will take a more exploratory approach. Instead of focussing on the capture and description of data, it will reflect on the significance of accurate self-assessment and its implications for teaching and learning in context.

The Ramifications of Accurate Self-assessment in Conjunction with Summative Assessment

While the sample in this study is too small to generalise from, the results from the analysis of the data agree with the general findings in the literature: Learners can self-assess to criteria with some accuracy. Should there be issues with the assessments the learners undertake, in terms of either construct validity or reliability of grading, it is possible that they may find themselves in classes they believe do not meet their needs, a situation which undermines the value of assessment. The spread of assessor scores in this project supports this view somewhat; all the participants were drawn from one class, and there was a spread of three full bands in the sample. Additionally, an over-preponderance of summative assessment and practice for high-stakes assessments has been shown to erode motivation in learners (Harlen and Crick, 2002). This leads me to hypothesise that inconsistencies between self- and summative assessments lead to motivational and other affective issues. If we accept that just over 70% of the learners (and 80% of learners who self-assessed) in this study are able to self-assess accurately, then perhaps they are also aware of the spread of competence in their class, thus undermining lower-achieving students' senses of self-efficacy and motivation. In a learning environment in Japan, where summative assessment is a dominant feature of many curricula, it may be reasonable to expect that a significant, but difficult to quantify, number of our learners question the validity of the conclusions drawn in the assessments they undertake, thus undermining their value. Additionally, if teachers are to be expected to assess their own learners, despite the strong evidence that the accuracy of such assessments is at least questionable, the role and scope of teacher assessment and self-assessment must be explored in far greater depth than is possible here.

Arguably the most pertinent question at this point, then, is the fundamental question any assessor, including teachers, must be able to answer satisfactorily: What is the purpose of assessment?

Japanese culture places great importance on tests, and attaining a high score on one is met with great approval. In that regard, therefore, tests in this country may be said to have some minimal kind of consequential validity (Messick, 1989) in that good performance in tests leads to social approval and encouragement. In the public school system, high-stakes tests such as entrance exams have the greatest consequences and thus frequent practice for summative assessments exert the greatest impact on the motivation of low-achieving students (Harlen and Crick, 2002). University entrance exams, for example, have far-reaching consequences in the lives of the children and young adults who take them. Once these tests have been taken, pass and fail grades awarded, and the candidate ensconced – by choice or not – in their new academic home, the meaning of frequent summative testing becomes rather less clear. In such circumstances, there appears to be little-to-no consequential validity for much of the summative

testing that takes place. To put it another way, if I visit a tailor to have my waist measured, I do it to ensure that my new trousers will fit properly, not for the dubious pleasure of being reminded that I am overweight. By the same token, testing that leads to the apportionment of scores which do not concur with learners' informal self-assessments and beliefs about their competence, and does not lead to any consequences for the learner, may be hypothesised to engender assessment fatigue and attendant issues with motivation, affect and classroom management. Additionally, Harlen and Crick (2002, p. 37) note that "[r]epeated practice tests showed [that] some students [were] all too well aware of what they could achieve and this led to very low views of their own capabilities." Such effects on self-esteem, derived from testing that potentially allocates a score rather than leading to feedback, undermine the usefulness of summative assessment in that low self-efficacy demotivates lower achieving learners, which in turn leads to lower self-efficacy. In other words, discussion of testing should not be limited to explorations of its administrative feasibility or its statistical validity; there is an ethical dimension to the discourse that cannot be ignored. As Mehrens (1990, p. 17) notes: "The bottom line in examining the results from an action is whether the positive consequences outweigh the negative." I submit that such a utilitarian concern is important for all educators, and that we must reflect on the purposes for which we deploy assessment and the consequences should we handle it wrongly – individually, institutionally or culturally.

Limitations

Primary among the issues in the study is the size and self-selecting nature of the sample: It is small and it is not random. Therefore, the findings cannot be generalised. Additionally, despite the amount of voice data generated by the interviews (11 interviews at about 14 minutes each, equalling some 150 minutes of interaction), the data generated is quantitative in nature and extremely limited. Such a small sample means that running even basic statistical analyses is of questionable value. However, the results are suggestive of further hypotheses and studies which may generate more statistically manipulable data. It is suggested that a larger-scale study involving a statistically significant number of students be undertaken.

In addition to statistical provisos, there are procedural issues. An important caveat to the Runnels paper cited (2013) is that it focuses on benchmarking the CEFR-J criteria and assumes that the self-assessments are accurate. The research reported here has a complementary goal – to explore the accuracy of self-assessment to criteria – and as such its limitations may be said to be complementary: The IELTS assessment criteria are well-benchmarked, but were adapted to render them comprehensible to learners and the reliability of this adaptation was not objectively assessed. Instead, I simplified the tool based on my understanding of the criteria and my knowledge and understanding of my students, and discussed possible changes in meaning with another suitably qualified practitioner. While I took steps to enable participant understanding of the items, it was not empirically ascertained whether meaning was faithfully preserved in the simplification process, nor whether the participants understood the criteria to mean the same as a trained and experienced assessor's understanding of them. Questions remain, therefore, over the reliability of the criteria used in the self-assessment.

A further proviso when considering the self-assessment data concerns familiarity with self-assessment protocols, which has been found to increase the accuracy of self-assessment (Muñoz and Albarez, 2007). As discussed above, the learners had opportunities to familiarise themselves with the criteria and to clarify any items they felt they did not understand. Nevertheless, self-assessment of language proficiency is not a prominent feature of the English program at the institution where the research took place. Therefore, although individual teachers working with the participants in the past may have had them perform self-assessments as part of previous English courses, the participants did not report this, and I therefore assume that they were not familiar with self-assessment practices. This lack of practice in self-assessment, combined with their varying proficiencies in English, may be viewed as a potential confounding factor in the data.

Conclusion

There is a very real possibility that current assessment practices in many educational contexts across the country – from compulsory education through tertiary institutions to corporate programs – are counterproductive, in that they work against the fostering of self-regulated learning and motivation, despite evidence that such feedback leads to improvements in learning outcomes (Clark, 2012). Current practices also potentially contribute to low motivation levels and other affective factors in our classrooms, possibly engendering assessment fatigue by clashing with the experiences and beliefs of the learners.

As is often the case with research projects involving people, the results raise more questions than they answer: Are the assessment tools we use reliable? Can inter-rater reliability be consistently demonstrated when the rater also teaches the students? Do our assessment tools allow us to make valid inferences regarding our learners' abilities in English? What roles should summative assessment, self-assessment and formative assessment play in curricula? Such questions require objective, though not unsympathetic, investigation, as they are not only empirical considerations, but also entry points to vital discussions surrounding the ethics of teaching, learning and testing: Our learners arguably have the right to understand both *how* and *why* they are being evaluated, with the expectation that there are fair and reasonable justifications for each and that the assessments they undertake will help them to develop.

References

Blanche, P., & Merino, B. (1989). Self-Assessment of Foreign-Language Skills: Implications for Teachers and Researchers. *Language Learning, 39* (3), 313-338.

Brown, R. (2004). EFL Learning Ability Self-Assessments of Japanese EFL Students. *Information and Communication Studies*, *30*(1)*,* pp. 1-6.

Bérešová, J. (2011). The impact of the Common European Framework of Reference on teaching and testing in Central and Eastern European contexts. *Synergies Europe*, *6*, 177-190.

Cambridge ESOL (2014). *Delta Module 1, Module 2, Module 3 – Handbook for tutors and candidates.* Retrieved on December 3rd, 2014 from http://www.cambridgeenglish.org/Images/181797-delta-handbook-2010.pdf

Cambridge ESOL (2010). *CELTA Syllabus and Assessment Guidelines*. Retrieved on December 3rd, 2014 from http://www.cambridgeenglish.org/images/21816-celta-syllbus.pdf

Cambridge IELTS 8 Student's Book with Answers: Official Examination Papers from University of Cambridge ESOL Examinations (IELTS Practice Tests). (2011). Cambridge: Cambridge University Press.

cefr-j.org (n.d.) CEFR-J (English Page). Retrieved on 20th of November 2014 from http://www.cefr-j.org

Clark, I. (2012). Formative Assessment: Assessment Is for Self-regulated Learning. *Educational Psychology Review, 24*(2), 205-249.

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

Council of Europe (2014). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (CEFR). Retrieved on 2nd of December, 2014 from http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

ETS (2013). *Can Do Guide Executive Summary: Listening and Reading*. Retrieved on 2nd December, 2014 from https://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_Can_Do.pdf

ETS (2004). *IBT/Next Generation TOEFL Tests Independent/Integrated Speaking Rubrics (Scoring Standards)*. Retrieved on 2nd December, 2014 from https://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

Gallagher, M. (2012). Self-Efficacy. In: *Encyclopedia of Human Behavior (Second Edition)* (Vol. 1, pp. 314-320). London: Academic Press.

Harlen, W. & Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning (EPPI-Centre Review, version 1.1). *Research Evidence in Education Library 1*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

IELTS (2014a) *Speaking Band Descriptors (Public Version)*. Retrieved on 2nd December, 2014 from http://www.ielts.org/pdf/Speaking Band descriptors_2014.pdf

IELTS (2014b) *Writing Task 1: Band Descriptors (Public Version).* Retrieved on 2nd December 2014 from http://www.ielts.org/PDF/Writing Band descriptors Task 1.pdf

IELTS (2014c) *Writing Task 2: Band Descriptors (Public Version)*. Retrieved on 2nd December, 2014 from http://www.ielts.org/PDF/Writing Band descriptors Task 2.pdf

Kuiken, F., & Veder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing*, *31*(3), 279-284.

Lowe, R. (2014). A Question of Qualifications: Making a Case for TESOL Diplomas. *OnCUE Journal, 6*(3), 54-61.

Mehrens, W. A. (1997). The Consequences of Consequential Validity. *Educational Measurement: Issues and Practice, 16,* 16–18.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York Macmillan

Muñoz, A. & Albarez, M. E. (2007). Students' Objectivity and Perception of Self Assessment in an EFL Classroom. *The Journal of Asian TEFL, 4*(2), 1-25

Runnels, J. (2013). Student ability, self-assessment, and teacher assessment on the CEFR-J's can-do statements. *The Language Teacher, 37*(5), 3-7.

Runnels, J. (2014). An Exploratory Reliability and Content Analysis of the CEFR-Japan's A-Level Can-Do Statements. *JALT Journal, 36*(1), 69-89.

Trinity College London (2007) *Licentiate Diploma in Teaching English to Speakers of Other Languages (LTCL Diploma TESOL) Validation Requirements, Syllabus and Bibliography for validated and prospective course providers*. Retrieved on December 3rd 2014 from http://www.trinitycollege.com/resource/?id=1776

Yamagishi, T., Hashimoto, H., Cook, K., Kiyonari, T., Shinada, M., Mifune, N., Inukai, K., Takagishi, H., Horita, Y. & Li, Y. (2012). Modesty in self-presentation: A comparison between the USA and Japan. *Asian Journal of Social Psychology, 15*(1), 60-68.

Yoshizawa, K. (2009). To what extent can self-assessment of language skills predict language proficiency of EFL learners in school context in Japan? 外国語教育研究（紀要）(*Foreign Language Education Research Bulletin*), *17*(1), 65-82.

Appendices

Appendix 1: The three parts of the IELTS speaking test

	The IELTS speaking test is in three sections, taking between 11 and 14 minutes to complete.
Part 1	Lasts up to 5 minutes. The assessor asks the candidate a series of questions regarding concrete everyday activities, preferences and interests and similar familiar topics.
Part 2	The candidate receives a task card, which specifies a topic and several subtopics. The candidate has up to one minute to prepare their response to the task. He/she then responds verbally for between one and two minutes. The tasks generally ask the candidate to describe a non-immediate, concrete experience, typically from their lives. The examiner may follow up with one or two questions on the same topic. This section lasts three to four minutes.
Part 3	This section is on the same general theme as part 2, although the questions are more abstract, including questions about society and culture, inviting the candidate to generalise and speculate. Part 3 lasts four to five minutes.

Appendix 2: The domains of the model of speaking competence in the IELTS speaking test, summarised from the IELTS Public Band Descriptors (Speaking) (IELTS, 2014a)

Fluency and coherence	How well-organised, well-signalled and easy-to-follow the speaker's English is. This includes discourse marking, use of pronouns and substitution, and clarity of expression. Negative criteria refer to repetition, pausing and their communicative impact.
Lexical resource	How well the candidate deploys vocabulary and paraphrases around gaps in lexical knowledge. This includes knowledge of idiomatic language and precision of expression, such as "excellent" in place of "very good". Consideration is given to the effect of errors and inappropriacies in word choice.
Grammatical range and accuracy	How well the interviewee uses the grammatical structures of English. The length and complexity of clauses and subordinate clauses is considered, including error density. Particularly important is the nature of the mistakes; whether they occur in simple sentences, or more complex utterances with several clauses and conjunctions.
Pronunciation	How well the candidate uses "acceptable pronunciation features" (IELTS, 2014a). The exact nature of these features is somewhat unclear. However, the lexical range criteria make specific reference to features of native speaker English such as idioms, and it is supposed that pronunciation features of inner-circle varieties of English might fit the "acceptable" criterion.

Appendix 3: The self-assessment tool

Fluency and Coherence	Lexical Resource	Grammatical Range and Accuracy	Pronunciation
I can take long turns without having to try hard. My meaning is always clear. Sometimes I pause to find the correct word or grammar. I can use many different signal words and linking words flexibly.	I have enough vocabulary to talk about a variety of topics easily. I can use some less common vocabulary and some idioms. I can use some vocabulary to show style. I know some collocations although sometimes I make mistakes.	I can use a mix of simple and complex sentences easily. I make few mistakes; sentences with no mistakes are common.	It is easy to understand me. I can use a wide range of contractions, stress and linked sounds to show exact meaning with few mistakes. I rarely make pronunciation mistakes.
I want to speak using long sentences and keep going, but sometimes the meaning is not clear because of repetition, self-correction or speech that is too slow. I can use several different signal words and linking words, but not always correctly	I can talk about familiar and unfamiliar topics for a long time. My meaning is usually clear, although sometimes I use the wrong words. I can usually paraphrase.	I can use a mix of simple and complex sentences, but not always correctly. I make mistakes often with longer sentences, but it is usually possible to understand the meaning.	I can use a range of stress, contractions and linked sounds but sometimes I make mistakes. I can use stress to show exact meaning, but not always.
I can usually keep going, but I use slow speech, repetition and self-correction I over use some signal words and linking words My simple speech is fluent, but more complex speech causes problems.	I can talk about familiar and unfamiliar topics, but unfamiliar topics are difficult. I try to paraphrase, but I don't always succeed.	I can usually use simple sentences correctly I can use a small mix of simple and complex grammar but these usually have mistakes and it can be difficult to understand me.	I try to use stress, but sometimes make mistakes. I can speak clearly, and I don't make many big pronunciation mistakes. I can usually be understood.
There are pauses when I answer I sometimes speak slowly with frequent repetition and self-correction I can link basic sentences, but I repeat the linking words often. Sometimes my meaning is difficult to understand	I can talk about familiar topics, but I can only communicate basic meaning about unfamiliar topics. I often choose the wrong words. I rarely try to paraphrase.	I can usually use correct simple grammar but longer sentences are rare. I make lots of mistakes. It is difficult to understand my meaning.	I can use some stress and contractions. I try to speak clearly, but I often make pronunciation mistakes. My pronunciation mistakes sometimes make it difficult to understand me.
I can speak with long pauses. I can link simple sentences sometimes. I can give simple responses. Often I can't say what I mean.	I can give personal information with simple vocabulary. I don't know enough words to talk about less common topics.	I can try to make sentences. Sometimes my grammar is good. I can repeat things I've memorised. I make lots of mistakes except when I have memorised the sentences.	I can sometimes use stress. I try to speak clearly, but I make pronunciation mistakes almost all the time. It is usually difficult to understand me.

Appendix 4: The interview questions

Part 1, set 1	1. Do you live in a house or an apartment? 2. What do you enjoy about where you live? 3. Would you like to live somewhere similar in the future?
Part 1, set 2	1. How well do you know the people who live next door to you? 2. How often do you see them? [Why?/Why not?] 3. What kinds of problems do people sometimes have with their neighbours? 4. How do you think neighbours can help each other?
Part 1, set 3	1. Which magazines and newspapers do you read? [Why?] 2. What kinds of article are you most interested in? [Why?] 3. Have you ever read a newspaper or magazine in a foreign language? [When?/Why?] 4. Do you think reading a newspaper or magazine in a foreign language is a good way to learn the language? [Why?/Why not?]
Part 2	Describe a restaurant that you enjoyed going to. You should say • where the restaurant was • why you chose this restaurant • what type of food you ate in this restaurant and explain why you enjoyed eating in this restaurant.
Part 3	Restaurants Why do you think people go to restaurants when they want to celebrate something? Which are more popular in your country, traditional restaurants or fast food restaurants? Why do you think that is? Some people say that food in an expensive restaurant is always better than food in a cheap restaurant – would you agree? Producing food Do you think there will be a greater choice of food available in shops in the future, or will there be less choice? What effects has modern technology had on the way food is produced? How important is it for a country to be able to produce all the food it needs, without importing any from other countries?