Kimura et al. Language Testing in Asia (2017) 7:3
DOI 10.1186/s40468-017-0035-2

Language Testing in Asia

CASE STUDY Open Access

Developing classroom language assessment benchmarks for Japanese teachers of English as a foreign language

Yuzo Kimura[1*], Yoshiyuki Nakata[2], Osamu Ikeno[3], Naoyuki Naganuma[4] and Stephen Andrews[5]

* Correspondence: ykimura@las.u-toyama.ac.j
[1]Faculty of Medicine, University of Toyama, Toyama, Japan
Full list of author information is available at the end of the article

Abstract

Introduction: Since the Japanese government has called for English to be taught through English (Ministry of Education, Culture, Sports, Science and Technology, 2003), there has been increasing interest in English language classrooms in upper secondary schools. The government has proposed English proficiency scores that Japanese teachers of English should attain, but has not yet specified clearly what English language use should be like in the classroom.

Case description: This case study describes the theoretical aspect of the development of a benchmark assessment for use in English language classrooms with Japanese teachers of English as a foreign language (EFL). The study first defines teacher language proficiency, the use of the foreign language in the classroom, with particular attention to teacher language awareness (Andrews, 2007) as a bridging mechanism between pedagogical content knowledge and the foreign language proficiency of Japanese teachers of EFL. This definition of teacher language proficiency is further elaborated in terms of the legitimate use of benchmark assessments in English language classrooms in Japan with a thorough literature review of L2 benchmark assessments in other parts of the world.

Discussion and Evaluation: The present case study examines the ideal assessor conditions after thorough review on assessor bias from the language testing/assessment research, while the development of assessment benchmarks for use in Japan is discussed based on a high-stakes benchmark assessment for teachers of EFL in Hong Kong. The proposed benchmarks acknowledge the complexity of classroom English use, thus employing four different scale types to accommodate the multifaceted characteristics of teacher language proficiency.

Conclusions: The current case study concludes with the exploration of tasks remaining for the future development of our benchmark assessments for use in ongoing professional development.

Keywords: Benchmark assessment, Classroom English use, Japanese teachers of EFL

Background

Since the late 1900s, the Ministry of Education, Culture, Sports, Science, and Technology (MEXT henceforth) in Japan has attempted to promote the English communication skills of high school students. This policy was partly realized by a series of decennial revisions of the Course of Study in 1989 and 1999, where the former introduced new oral

communication subjects (Ministry of Education, Culture, Sports, Science and Technology MEXT 1989), while the latter called for the development of "practical communication abilities" (Ministry of Education, Culture, Sports, Science and Technology MEXT 1999, p. 3) as part of fundamental reforms. In order to make this communication-oriented policy vision more stable, however, MEXT further announced the 5-year Action Plan to Cultivate "Japanese with English Abilities" (Ministry of Education, Culture, Sports, Science and Technology MEXT 2003, p.2), through which a number of reforms have been proposed. This includes active collaborations with assistant language teachers (ALTs henceforth) from the JET Program(me)[1] to create more elaborate communicative classes based on the degree of learners' academic achievement, promotion of the 100 Super English High School Project, and supporting English activities at elementary school which eventually led to the introduction of a new compulsory curriculum for elementary schools called 'Foreign Language Activities' in 2011 (Ministry of Education, Culture, Sports, Science and Technology MEXT 2011).

Amid these reforms were two more crucial proposals in terms of English language teaching in Japanese high school; the definition of English proficiency for Japanese teachers of EFL for upper secondary schools using three different English proficiency test scales and the declaration that, "The majority of an English class will be conducted in English" (Ministry of Education, Culture, Sports, Science and Technology MEXT 2003, p.2). While the former was the first attempt in history to specify the English proficiency that Japanese teachers of EFL are expected to hold[2], the latter became the backbone of MEXT's 2009 version of the Course of Study.

In the new 2009 Course of Study for upper secondary schools, a statement was incorporated about the medium of instruction in English classes, specifying that "classes, in principle, should be conducted in English in order to enhance the opportunities for students to be exposed to English, transforming classes into real communication scenes" (Ministry of Education, Culture, Sports, Science and Technology MEXT 2009, p. 7). Since this statement's release, conducting English lessons in English has been hotly debated (Clark, 2009, February 5; Hato, 2005). However, there have been no official guidelines so far setting out what classroom language is necessary for Japanese teachers of EFL and thus, it remains an urgent task to both clarify the nature of the English proficiency Japanese teachers of EFL should attain and to explore ways to support those teachers' ongoing language development. To address this issue, this study proposes a benchmark assessment of English teachers' classroom English use consisting of four independent but mutually related scales in the hope that teachers can use these scales for self-reflective professional development.

While an earlier, shorter manuscript describing the development of these scales has been published elsewhere (Nakata, IKeno, Naganuma, Kimura, & Andrews 2012), this more theoretical paper is concerned with the detailed step-by-step process of the development of this benchmark assessment. This is done by first defining the teacher language proficiency required by Japanese teachers of EFL. Then, we review various foreign language assessment scales from different sociocultural contexts, along with some important issues related to rater performance in evaluating spoken language. Having described the background to the study, we explain and examine the development of our classroom language benchmark assessment, which is followed by a discussion of potential implications for future research.

The definition of teacher language proficiency for Japanese teachers of EFL

Defining the language proficiency required for teachers of EFL is not an easy task. This is particularly the case since, as Pasternak and Bailey (2004) rightly explain, "Whether or not a teacher is proficient depends on how we define this multifaceted construct" (p. 163).

The challenges in determining the language proficiency necessary for Japanese teachers of EFL can be associated with the broader challenge of unpacking the relationship between the knowledge required for teaching and their English language proficiency. Bachman (1990) describes "communicative language ability" (p. 84) in terms of language competence, strategic competence and psychophysiological mechanisms. While this definition is fundamental to many language proficiency tests, including those cited in Ministry of Education, Culture, Sports, Science and Technology (MEXT) (2003), Shulman's (1987) knowledge base is another way to establish a baseline of knowledge for teachers. Shulman (1987) summarizes seven categories of knowledge necessary for teaching, of which pedagogical content knowledge is of special interest as it "represents the blending of content and pedagogy into an understanding of how particular topics, problems, or issues are organized, represented, and adapted to the diverse interests and abilities of learners, and presented for instruction" (Shulman, 1987, p.8). It seems fair to conclude then that for Japanese teachers of EFL, teacher language proficiency comprises the combination of both individual foreign language proficiency and pedagogical content knowledge. The remaining discussion turns to addressing how these two areas of language use are interconnected.

Andrews (2007) has sought to investigate the nature of pedagogical content knowledge as it relates to teachers of foreign languages, using the term teacher language awareness (TLA henceforth) to refer to one subset of knowledge that is arguably unique to the teacher of a second or foreign language (L2 henceforth), which interacts with other subset components and blends with them in the act of expert L2 teaching. In this regard, TLA is seen as a bridge between pedagogical content knowledge and foreign language proficiency, since it involves reflections on both and the mediation of the former through the latter, as well as awareness of the learners and their developing L2.

For foreign language learners, classroom target language input originates from three main direct or 'unfiltered' output sources; language in materials (particularly when students engage in self-study), language produced by other learners (when they interact with each other in L2) and language produced by the teacher. All of these input sources also reach them as 'filtered' via the teacher's classroom language. The teacher, for example, can modify textbook language to make it more easily comprehensible. When the learners talk with each other, teacher feedback often becomes an additional source of input in the target language and the learners' original output may be 'filtered' for them by the teacher (Andrews, 2007). Therefore teachers, if they are language aware, are able to pay careful attention to their TLA 'filter', which potentially results in substantial classroom learning benefits. Seen from another perspective, it may be reasonable to say that the classroom language use of teachers with their learners can be enhanced largely through raising their TLA. This view is an example of a concrete conceptualization of Krashen's 'input hypothesis' (Krashen, 1985), wherein comprehensive input can lead to learner L2 acquisition. If teachers want their classrooms to be rich in L2 acquisition, they need to pay careful attention to their classroom language

use so that the L2 language environment stays within the range of comprehensible input, which can be accomplished through enhancing TLA. In addition, while most empirical studies investigating the positive influence of teaching skills on learners tend to examine skills such as form-focused instruction in grammar teaching (Ellis, 2002) or tasks in the L2 classroom (Ellis, 2003), TLA itself as a mechanism for enhancing learning outcomes remains relatively unexplored. It is here that our interest in how to raise TLA lies, and as a result how to enhance teacher classroom language proficiency.

TLA can also play a significant role in light of other SLA disciplines. From a neo-Vygotskian sociocultural standpoint (Lantolf & Thorne, 2006), for example, TLA has the potential to influence the decisions teachers make as to whether to provide scaffolding or not (Andrews, 2007). Teacher utterances are also crucial as a subject of "ventriloquation" (Wertsch, 1991, p.59); the more learners are cared for by a language aware teacher who can use his/her classroom language in an authentic way, or as people use it in the real world outside of school, the more chances learners have to populate and appropriate their teacher's voice, through which they can realize authentic language use in the classroom.

The issue that still remains then is how we can raise TLA in order to sustain both foreign language proficiency and pedagogical content knowledge hand in hand. It is here that the classroom language benchmark assessment comes into play as a tool for stimulating TLA, thereby enhancing and developing individual teacher language proficiency. This is accomplished through the benchmark assessment providing current information about where a given teacher's teacher language proficiency is and suggesting directions for potential future improvement. In the next section, we review benchmark assessments for second and foreign language teachers in order to locate our benchmark assessment relative to measures that precede it.

Benchmark assessments for L2 teachers

When language proficiency is assessed, it is necessary to have a standard on which to base the assessment. Although standards may be described differently, such as through bandscales, benchmarks or curriculum frameworks, they should provide precise and comprehensive descriptions of the knowledge and abilities necessary at different levels of proficiency. While there are benchmark assessments for adult learners (e.g., *the Canadian Language Benchmarks Assessment* (Centre for Canadian Language Benchmarks, 2000); *the Common European Framework of Reference for Language* (CEFR henceforth) (Council of Europe, 2001), the need for benchmarking or minimum standards in the teaching profession started to be discussed in the 1970s and to emerge more explicitly in the 1980s. Below we review three such cases, two from Europe and one from the Far East.

In Europe, the Council of Europe's *European Centre for Modern Languages* (ECML) is responsible for the *European Portfolio for Student Teachers of Language* (EPOSTL). The rationale for the EPOSTL came partially from the CEFR, a guideline describing a common basis of foreign language teaching/learning guidelines across Europe, in that both try to describe competencies in the form of can-do descriptors (Newby, 2007). However, unlike the CEFR, the EPOSTL does not use any numerical scale since the ECML believed didactic knowledge to be unsuited to quantification (Council of Europe,

2007). Built from "a bottom-up view" (Newby, 2007, p. 24) through feedback from student teachers and teacher educators, the EPOSTL consists of nine sections, three of which are working sections that require student teachers to perform tasks: the *Personal Statement*, the *Self-Assessment*, and the *Dossier*. The *Self-Assessment* section is the heart of the EPOSTL and has seven general categories, each of which is further divided into three to seven sub-topics, providing a total of 32 areas and 195 can-do descriptors. In terms of Teacher Language Proficiency, however, there is only one sub-category; *Classroom Language* in the *Conducting a Lesson* general category. There are only six semi open-ended descriptors (out of 195) in this section to describe classroom language use in open-ended description manner. We believe these descriptors are insufficient for the current purpose of assessment of teacher classroom language use because, first, the number of descriptors itself is too small to comprehensively assess language use in a lesson; second, self-descriptions of these six descriptors may be able to capture subjective impressions of pre-service trainees' classroom language use, but do not necessarily capture the multifaceted characteristics of classroom language use; and third, a lack of a numerical assessment scale may prevent independent evaluation by other assessors.

The second example of benchmarks for teachers is the Evaluation & Accreditation of Quality Language Service (EAQUALS)'s *The Profiling Grid for Language Teachers.* Inspired by CEFR, this benchmark assessment grid consists of a) a set of reference levels as three *Stages* and b) a one-page open-ended inventory describing the professional skills required of language teachers. The three *Stages* are "Basic", "Independent" and "Proficient" and reflect the three levels of the CEFR, whereas the open-ended inventory includes four categories, "Language", "Qualifications", "Core Competencies", and "Complementary skills". The "Language" category has "Language Proficiency" and "Language Awareness", both of which are relevant to our interest in classroom language use for our assessment tool. The largest gap between their and our benchmark assessment is the fact that the grid originally meant to describe the profile of practicing language teachers at EAQUALS member schools as part of preparation for inspection by EAQUALS (Rossner, 2009, February). As such, the Grid provides a framework for stages of development of language teachers' pedagogy rather than for assessing teaching quality (North, 2009, February).

The third example is Hong Kong's benchmark assessment for EFL teachers: the Language Proficiency Assessment for Teachers (English language) (LPATE henceforth). In 1995, the Hong Kong Government passed two recommendations to address concerns over the perception of falling language standards. They were (a) the initiation of benchmark qualifications for all teachers of English, and (b) the specification of minimum language proficiency standards for teachers to attain professional qualification. With these recommendations, in 2001, the first live LPATE was administered in three assessment areas; (a) language ability, (b) subject content knowledge and (c) pedagogical content knowledge (Coniam & Falvey, 2001).

After several revisions, the current 2007 version of the LPATE has expanded to five different areas of assessment of EFL teachers' abilities; (a) a set of three formal 'pen-and-paper' assessments for reading, writing and listening, (b) an additional speaking assessment [3]and (c) a Classroom Language Assessment (CLA hereafter). All five of these assessments are mandatory for all teachers of English in Hong Kong, representing a 'high-stakes' assessment for them because it affects their careers. Also, unlike other benchmarking assessments for teaching professionals which focus mainly on

assessing subject matter knowledge or the foreign language proficiency of applicants, the CLA component of the LPATE observes the English language use of EFL teachers in an actual classroom. In the CLA, English teachers are assessed twice by two different assessors in a live lesson on four different constructs; (a) grammar and lexical accuracy and range, (b) pronunciation, stress and intonation, (c) the language of classroom interaction and (d) the language of classroom instruction. These four constructs were originally devised by the Working Party for the CLA in Hong Kong through inductive categorization after watching 20 classroom videos over six meetings (Coniam & Falvey, 1999). Each construct is scored on a five-level scale with the mid-point (level 3) being 'at the benchmark level,' which is seen as the minimum standard that all EFL teachers in Hong Kong need to achieve. Attainment of the overall benchmark requires at least '2.5' or above on any one construct and '3' or above on all other constructs (Education Bureau Government of the Hong Kong Special Administrative Region 2007).

In his study, (Nakata, 2010) verified the applicability of the LPATE's CLA benchmark as a tool for TLA development of Japanese teachers of EFL in Japan. Eight masters students in his graduate course, five were in-service whereas three were pre-service Japanese teachers of EFL, were participated in the assessment of their peers' classroom English use in microteaching sessions. A post-hoc 6-item questionnaire was carried out five months after the course, asking them such questions as to what extent they felt the CLA benchmark can improve teacher English proficiency, or to what extent the assessment of teacher classroom English can be meaningful. Nakata (2010) conducted a follow-up study with six of the same participants one year later; assessing classroom English use using a revised CLA sheet and a revised 7-item post-hoc questionnaire administered five months after the follow-up study course. The data from these two questionnaires were collected included both 5-point *Likert* scales and open-ended written feedback. The results clearly suggested that the classroom English observation program conducted with the CLA benchmark enhanced trainees' awareness of classroom English use and showed a strong potential to improve their TLA. There was also some important feedback from the participants regarding how to more finely tune future revisions of the CLA for secondary school contexts in Japan. These include requests for more detailed subscale descriptors of *interaction* and *instruction*. As such, direct application of this CLA benchmark based on the LPATE to the Japanese context was determined to be potentially inappropriate and likely insufficient. However, the research also suggests the LPATE's CLA can be an example that our benchmark assessment can be safely based on and developed from. With this suggestion in mind, the design and the development of our own version of a classroom language assessment benchmark is described in the following section.

Discussion and evaluation

Designing classroom language assessment benchmarks

As part of making our language assessment benchmark salient to English language teaching in Japan, there are several crucial differences between it and LPATE's CLA. First, the LPATE is a high-stakes assessment for all teachers of EFL in Hong Kong, whereas our assessment is low-stakes in nature. The rationale for our benchmark assessment to be low-stakes is three-fold. First, currently in Japan it is unrealistic and impractical to conduct

this type of assessment in a top-down fashion, as there is no official policy regarding establishing classroom English use assessment for upper secondary schools at a national or regional scale. Secondly, we believe assessing classroom English use, although crucial for English teachers' professional development, can be and should be conducted on a voluntary rather than mandatory basis, as it requires a certain level of courage for teacher assesses to confront their own English proficiency, and often involves delicate and painful emotions. Finally, developing a multifaceted benchmark assessment for classroom language use requires a long-term cyclical process of careful item selection, implementation of field-testing and follow-up revisions. We believe low-stakes bottom-up assessment allows for such careful and steady development.

The difference between high and low stakes assessment has a crucial impact on all aspects of benchmark development, including objectives, targets and the number of scales. The main aim of administering the LPATE as a high-stakes benchmark in Hong Kong is to discriminate between the classroom language use of EFL teachers and to maintain their English teaching quality at or above an established minimum level. To accomplish this, the benchmarks are focused on certain limited scales, a single set of four constructs in this case, for practicality of administration. Our benchmark assessment, on the other hand, is intended to be used for professional development on a voluntary basis at a smaller scale, to encourage English teachers to be more aware of their classroom English use, for example in post-lesson observation conferences in schools and graduate programs or in-service teacher seminars at pre-service or in-service training sessions. In such situations only a single set of four construct scales would be inadequate to describe a given EFL teacher's multifaceted teacher language proficiency. Second, while the five-level rating method with Level 3 being 'acceptable' is used in LPATE's CLA, our benchmark assessment employs a four-level rating method with Level 2 as the 'acceptable' benchmark. This choice was made because in conducting teacher-friendly benchmark assessment within a certain limited time frame with teacher development in mind, it is believed that the number of rating levels should be as manageable as possible, while, at the same time, the 'acceptable' benchmark level should not be set too high. The description of Level 1 in our benchmarks is also labeled as 'not yet acceptable', rather than being worded more negatively, in the hope that those who are assessed will be encouraged to persevere in their efforts to attain a higher level of competence. Third, since the main objective of the LPATE is to assess the English proficiency of EFL teachers, it pays less attention to the function of TLA, while our benchmark assessment is intended to facilitate EFL teachers' TLA and encourage them to pay more attention to their English use in the classroom, thereby providing better learning environments for their students. Keeping these contrasts between the current benchmark assessment and LPATE's CLA in mind, we turn to describing the development of each of the four scales in our benchmark assessment.

Developing the four scales for the classroom language assessment benchmark

Because of the multifaceted nature of teacher language proficiency, we believe classroom English use similarly needs to be conceptualized in a multifaceted way so that Japanese teachers of EFL, pre-service and in-service, can concentrate on different aspects of their development needs at different times. In this regard, developing four

complimentary benchmark scales offers teachers multiple options for focusing on assessing and improving different aspects of their teaching and makes administration easier, as the four scales can be used independently or in combination, depending on available teacher development time and resources. Each scale has its own distinctive features which allow for shedding light on particular aspects of classroom English use, thus enabling Japanese teachers of EFL to be more aware of their classroom English at different developmental stages. This is not to claim that our scales are better than other measures by simply employing multiple dimensions of assessment, but rather to state that as the objective of our benchmark is professional development, it seems natural that teachers' developmental issues with their classroom language use will differ from person to person, and even from instance to instance with the same individual. Therefore our benchmark, which seeks to facilitate teacher awareness of their classroom English use, should likewise account for such potential variation.

The four separate but interrelated scales in our benchmark assessment are: (a) an Integrated Diagnostic Scale (see Additional file 1: Appendix 1): a rubric for external assessment of global use of English in the classroom, (b) Reflective Analytic Scales (see Additional file 1: Appendix 2): scales for self-reflective use to self-assess classroom English use, (c) Function-specific Scales (see Additional file 1: Appendix 3): self-reflective scales focusing on the functional aspects of EFL teacher English use in the classroom and (d) Task-specific Scales (see Additional file 1: Appendix 4), which consist of rubrics developed to assess classroom English use in relation to various tasks frequently employed in the context of high school EFL classrooms in Japan. These four scales are all intended to raise the TLA of EFL teachers and thus to encourage them to pay further attention to enhancing their teacher language proficiency. Providing multiple scales in the way we have here allows for accommodating individual needs with greater ease than through one larger, more general universal measure.

The issue of how EFL teachers' classroom English use is scored should also be addressed in developing these four scales, including consideration of assessor bias. In language testing/assessment studies, the issue of scoring as a subjective assessment of spoken language (Davis, 2016; Sato, 2011) has been raised, and there seem to be two important factors that may influence the performance of our assessors; their scoring training and the so-called halo effect (Throndike 1920). The former concerns to what extent the training of raters contributes to the consistency of their scoring. The latter refers to raters' assessment of one dimension of performance influencing their assessments of other dimensions of performance (Throndike 1920).

In the field of language testing research it is generally believed that rater training is necessary to maintain the reliability and validity of language performance tests (Fulcher, 2003). Some empirical studies do support this view, suggesting higher inter-rater reliability and agreement after training (Shohamy et al. 1992; Weigle, 1994, 1998). Others find training results in considerable variation in rater severity and scoring criteria (Lumley & McNamara, 1995; Orr, 2002; Papajohn, 2002). With such inconsistent research evidence, regarding the scoring of the TOEFL iBT Speaking Test by 20 native teachers of English, Davis (2016) confirmed rater training led to modest improvements in inter-rater reliability and agreement whereas it had little impact on rater consistency or severity.

The second element of concern is the "halo" (Throndike 1920, p.28) effect; when a rater's judgment of a single element of evaluation influences other elements in the assessment. For example, Yorozuya and Oller (1980) found it between two different rating procedures. In a five-time listening scoring experiment they conducted with 15 native speakers of English raters, interviews with 10 foreign students were evaluated in two ways; rating only one out of four scaling constructs of English speech (i.e., grammar, vocabulary, pronunciation and fluency) individually on four independent consecutive hearings versus rating all four scaling constructs on the remaining one listening. The halo effect was revealed on the single hearing occasion, and it tended to reduce the reliability of the raters' scores. Yorozuya and Oller (1980) speculate that the raters may have been biased by their previous scoring. Bechger et al. (2010) suggest a practical way to avoid halo effects by assigning "raters at random to combinations of examinees and assignments" (Bechger et al. 2010, pp. 616–617). The reliability of CLA in the current 2007 version of the LPATE has been maintained by either double marking or assessment on two separate occasions by assessors from the Hong Kong Government's Education Bureau (Coniam & Falvey, 2013).

To put these findings into the context of raters for our benchmark assessments discussed here, it would be fair to conclude that (a) the number of assessors should be at least two, up to as many as practical for the assessment occasion, perhaps with three to five assessors as the most reasonable number; (b) the assessors should include at least one who is unfamiliar with the assessee; and finally, (c) all assessors should complete a training/practice process before the actual assessment.

The integrated diagnostic scale

The Integrated Diagnostic Scale was developed based on the LPATE's CLA with the intention of assessing overall levels of teacher language proficiency in a complete lesson. LPATE's CLA has a single set of four constructs, (a) grammar and lexical accuracy and range, (b) pronunciation, stress and intonation, (c) the language of interaction and (d) the language of instruction. These four constructs can be further summarized into two categories: formal elements which define an English language teachers' English ability; and functional realizations of a teacher's formal English ability (Coniam & Falvey, 1999). Therefore in LPATE's CLA, two theoretically different construct aspects are incorporated into one assessment scale. In contrast, the Integrated Diagnostic Scale in our benchmark assessment has five constructs; (a) grammar (accuracy & variety: to what extent those who are assessed can use a wide range of English grammar accurately), (b) vocabulary (appropriateness & variety: to what extent English vocabulary selections of those who are assessed are appropriate and rich in variety), (c) pronunciation (accuracy & naturalness: to what extent those who are assessed can pronounce English accurately and naturally), (d) instruction & explanation (efficiency & clarity; to what extent instructions of those who are assessed are efficient and clear) and finally, (e) interaction with students (smoothness: to what extent those who are assessed smoothly interact with students), with no linguistically hierarchical difference. In the Integrated Diagnostic Scale, grammar and vocabulary are independently scaled. This is because we expect our benchmark assessment to be as informative as possible and if grammar and vocabulary are assessed together, it could lose the ability to differentiate between the two and thus may decrease its potential to improve their TLA.

With this scale, the attainment of an overall level of '4' requires '4' in at least four of the five constructs and '3' in only one construct (See Case A in Table 1). Likewise, a candidate is judged at level 3 either when there is more than one construct at '3' while the rest are '4' (Case B in Table 1), or there is one '2' while the rest are all more than '2' (Cases C and D in Table 1). Finally, a candidate is judged at level 2 when there are two or more constructs at '2' regardless of any upper levels for the other constructs. If the candidate has a single '1' in any construct, they are not considered to be at the benchmark level.

Reflective analytic scales

Using the same scale constructs as the Integrated Diagnostic Scale described above, the Reflective Analytic Scales are designed to guide EFL teachers' self-reflections. Unlike the Integrated Diagnostic Scale which is for use by assessors who have observed a complete lesson, the Reflective Analytic Scales are a tool for self-assessment through answering 'can-do' descriptors immediately after a class. We believe this unique introspection/retrospection feature of the Reflective Analytic Scales will enable those who are assessed to express more qualitative impressions regarding their classes, thereby complementing data obtained from the Integrated Diagnostic Scale. This is particularly beneficial for use with high school Japanese teachers of EFL as they tend to prefer to keep their English proficiency levels private, and yet they are fully aware of the need for professional development.

Function-specific scales

The functional aspect of classroom language use is assessed separately from the Integrated Diagnostic Scale through Function-Specific Scales covering six different functions: (a) elicitation, (b) facilitation, (c) clarification request, (d) recasts, (e) comments and (f) assessment. These functions are all interactional in nature and were devised through deductive analysis of descriptors from the CLA, looking for interactional language useful for the language classroom. First, two superordinate functions were identified (elicitation and feedback). Next, further consideration revealed that under the elicitation function, three more subordinate functions were necessary (elicitation, facilitation and clarification requests), while under the feedback function, another three subordinate functions were included (recast, comment and assessment). These 'can-do' descriptive scales are intended to be used by EFL teachers self-reflectively to check the degree to which they can perform the target functions effectively.

Table 1 Sample results of assessment of overall assessed level of one teacher's teacher language proficiency by five assessors

Assessors Scale constructs	A	B	C	D	E
Grammar	4	4	4	4	4
Vocabulary	4	4	4	4	4
Pronunciation	4	4	4	3	3
Instruction & Explanation	4	3	3	3	2
Interaction with students	3	3	2	2	2
Overall Level of	4	3	3	3	2

Task-specific scales

The Task-Specific Scales were developed to assess English teachers' classroom language use when dealing with tasks which they are likely to be familiar with in the Japanese EFL classroom context. The candidate teachers choose tasks which they think are relevant to their own lessons because the target tasks are determined by the pedagogical approach each teacher adopts. These tasks are evaluated by assessors through lesson observation (either live or videotaped). Therefore, the Task-Specific Scales can be understood as primarily embodying instructional aspects of teacher language proficiency, specific to Japanese EFL classroom contexts, and far more independent and detailed than the equivalent constructs for 'The Language of Instruction', one of a single set of four scales in the CLA in the LPATE which describes the way in which a teacher interacts with students.

The Task-Specific Scales were developed in a primarily inductive manner. First, various tasks were selected and then their assessment constructs and level descriptors were drafted. Second, these first drafts were scrutinized by several high school EFL teacher collaborators. Third, some revisions were made before reaching the current version, which includes ten tasks. They are (a) oral reading practice, (b) oral introduction of the target passage content, (c) introduction of the target grammar, (d) provision of relevant background knowledge, (e) comprehension check, (f) interaction with ALTs in team teaching, (g) oral exchange with students as a whole class, (h) instruction of words and idioms, (i) presentation of a model speech, and finally, (j) presentation of an oral summary (see Additional file 1: Appendix 4 for a sample of an oral reading practice assessment).

The relationship between the four scales and implications for their use

Table 2 summarizes the relationship between the four scales and Fig. 1 graphically illustrates the four scales along with their distinctive features in two dimensions. It also shows how reflection is an essential component of all fours scales and how these scales can be used to assess a lesson. In terms of the horizontal dimension in Fig. 1, for example, the Integrated Diagnostic Scale and the Task-Specific Scales both use observation-based reflection with 'be-doing' descriptors determined by the assessors observing the class, whereas the Reflective Analytic Scales and the Function-Specific Scales use self-assessed, introspection-based reflection with 'can-do' descriptors. In terms of the vertical dimension, the Integrated Diagnostic Scale and the Reflective Analytic Scales can be applied to an entire lesson observation while the Function-Specific Scales and the Task-Specific Scales are for assessing only part of a lesson. Two related issues are discussed here.

Table 2 Relationship between the four scales

	Integrated diagnostic scale	Reflective analytic scales	Function-specific scales	Task-specific scales
Main objective	External evaluation & assessment	Reflection	Reflection	Lesson analysis
High Reliability	Required	Desirable	Desirable	Desirable
Users	Other assessors	The teacher who conducted the lesson	The teacher who conducted the lesson	Other assessors
Unit of evaluation/ analysis	Entire lesson	Entire lesson	Language function	Task (selected by teacher)
Based on lesson observation?	YES	NO	NO (Possible)	YES

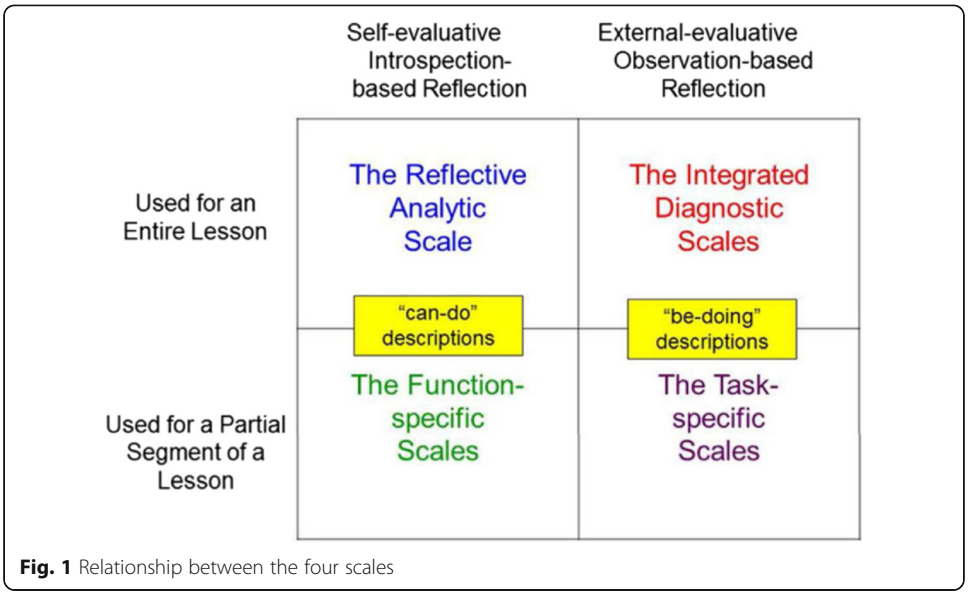

Fig. 1 Relationship between the four scales

First, because of their observation-based reflective nature, both the Integrated Diagnostic Scale and the Task-Specific Scales can be used to interpret language proficiency of Japanese teachers of EFL based on observable classroom phenomena in a lesson. This reflects the importance of teacher language proficiency as it relates to interaction with students in conjunction with the extent to which students are involved in and interact with teacher instruction in English. Therefore, if teachers with considerable English proficiency talk to their students without eliciting any response, the rating of their performance for *Interaction with students* should reflect that lack of interaction with the students, however fluent the teacher's English may be.

Secondly, because of these independent but mutually-related scales, various applications can be proposed. At a school-based teacher training session for a young novice teacher, for example, experienced colleagues can use the Integrated Diagnostic Scale to evaluate an entire class or the Task-Specific Scales to make a focused evaluation of a particular task performance, while the teacher who is assessed can use the Reflective Analytic Scales to capture their own impression of their teaching immediately after a class, or the Function-Specific Scales to check and evaluate the effectiveness of target functions in the lesson by viewing a video of their teaching practice.

Conclusions

In this paper, we have described the development of classroom language benchmark assessments for Japanese teachers of EFL with an emphasis on the theoretical rationale for their design. The current pilot version of the assessment scales presented here represents an initial step toward a more complete and practical assessment of classroom English use by Japanese teachers of EFL. Care will therefore need to be taken in their implementation to evaluate their effectiveness and ensure their future development.

At this stage, there are three specific issues that must be addressed going forward. First, the constructs of teacher language proficiency and each scale descriptor must be further refined in terms of their theoretical rationale. Particular attention should be paid in identification and addition of key characteristics for each descriptor. Secondly,

regarding practicality of assessment, the adjustment of scale levels and checking of the validity of the scales remains incomplete. Unlike other benchmark assessments, our version has four different scales, and this may require more complex and time-consuming procedures for assessors and assessees. Smooth administration is crucial for the dissemination of our benchmark assessment to the relevant educational authorities. Last, but not least, further field-testing will be indispensable to making the current assessment more practical for future use.

The above refinements are crucial to our benchmark assessment being readily adopted by teachers and school authorities for utilization for its intended purposes, in pre-service and in-service training sessions or in post-observation discussions, as a basis for the professional development of Japanese EFL teachers.

Endnotes

[1]Assistant Language Teachers (ALTs) refers to non-Japanese teachers hired to team-teach English along with Japanese teachers of English. They include teachers hired by the Japanese government through the Japan Exchange Teacher (JET) program(me). ALTs are also hired via local boards of education, both directly and indirectly through contracts with private outsourcing companies.

[2]STEP pre-first level, TOEFL 550, TOEIC 730 or over. The Society for Testing English Proficiency (STEP) has from grade 1 (equivalent to CEFR C1) to grade 5 (CEFR A1). The Grade pre-first level is equivalent to CEFR B2.

[3]The speaking assessment in LPATE was revised in June 2010.

Additional file

Additional file 1: Appendix 1. Integrated Diagnostic Scale. **Appendix 2.** Reflective Analytic Scales. **Appendix 3.** Function-specific Scales. **Appendix 4.** Task-specific Scales: Oral Reading Practice. (DOCX 31 kb)

Acknowledgement
This work was supported by JSPSKAKENHI, the Grant-in-Aid for Scientific Research (C) (Grant Numbers 22530969, 26381199) (Project Leader: Yoshiyuki Nakata).

Authors' contributions
YK, YN, IO and NN discussed and developed the proposed benchmarks, while SA provided some advices based on his experiences of LPATE in Hong Kong. YK drafted the manuscript: the other authors provided insightful comments, YN as a research leader and SA as an expert of this field in particular. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Faculty of Medicine, University of Toyama, Toyama, Japan. [2]Faculty of Global Communications, Doshisha University, Kyoto, Japan. [3]Faculty of Education, Ehime University, Tatsuyama, Japan. [4]International Education Center, Tokai University, Hiratsuka, Japan. [5]Faculty of Education, the University of Hong Kong, Hong Kong, China.

Received: 27 September 2016 Accepted: 1 February 2017
Published online: 13 February 2017

References

Andrew, S. (2007) *Teacher language awareness*. Cambridge: Cambridge University Press.
Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
Bechger, T. M., Maris, G., & Hsiao, Y. P. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement, 34*(8), 607–619. doi:10.1177/0146621610367897.
Centre for Canadian Language Benchmarks. (2000). *Canadian language benchmarks 2000 English as a second language for adults*. Ottawa: Centre for Canadian Language Benchmarks.
Clark, G. (2009). What's wrong with the way English is taught in Japan? *Japan Times*. Retrieved from http://www.japantimes.co.jp/opinion/2009/02/05/commentary/whats-wrong-with-the-way-english-is-taught-in-japan/#.Wjbb0fmLSUk. Accessed 5 Feb 2017.

Coniam, D., & Falvey, P. (1999). The English language benchmarking initiative: A validation study of the Classroom Language Assessment component. *Asia Pacific Journal of Language Education, 2*(2), 1–35.
Coniam, D., & Falvey, P. (2001). Awarding passes in the language proficiency assessment of English language teachers: Different methods – varying outcomes. *Education Journal, 29*(2), 23–35.
Coniam, D., & Falvey, P. (2013). Ten years on: The Hong Kong language proficiency assessment for teachers of English (LPATE). *Language Testing, 30*(1), 147–155. doi:10.1177/0265532212459485.
Council of Europe. (2001). *Common European framework of reference for languages: Learning teaching, assessment*. Cambridge: Cambridge University Press.
Council of Europe. (2007). *European portfolio for student teachers of languages: A reflection tool for language teacher education* [Adobe Digital Editions version]. Retrieved from http://www.ecml.at/epostl. Accessed 5 Feb 2017.
Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135. doi:10.1177/0265532215582282.
Education Bureau Government of the Hong Kong Special Administrative Region. (2007). *Language proficiency assessment for teachers (English language)*. Hong Kong: Education Bureau Government of the Hong Kong Special Administrative Region.
Ellis, R. (2002). The place of grammar instruction in the second/foreign language curriculum. In E. Hinkel & S. Fotos (Eds.), *New perspectives on grammar teaching I second language classrooms* (pp. 17–34). Mahwah, NJ: Lawrence Erlbaum.
Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.
Fulcher, G. (2003). *Testing second language speaking*. Harlow, UK: Longman.
Hato, Y. (2005). Problems in top-down goal setting in second language education: A case study of the "Action Plan to Cultivate 'Japanese with English Abilities". *JALT Journal, 27*(1), 33–52.
Krashen, S. (1985). *The input hypothesis: Issues and implications*. London: Longman.
Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford: Oxford University Press.
Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71. https://doi.org/10.1177/026553229501200104.
Ministry of Education, Culture, Sports, Science and Technology (MEXT). (1989). *The Course of Study for Upper Secondary Schools*. Tokyo: MEXT.
Ministry of Education, Culture, Sports, Science and Technology (MEXT). (1999). *The Course of Study for Upper Secondary Schools*. Tokyo: MEXT.
Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2003). *'Eigoga tsukaeru nihonjin' no ikusei no tame no kodo keikakuno sakutei ni tsuite* [*Regarding the establishment of an action plan to cultivate 'Japanese with English abilities'*]. Retrieved from the MEXT website: http://warp.da.ndl.go.jp/info:ndljp/pid/242299/www.mext.go.jp/b_menu/houdou/15/03/030318a.htm. Accessed 5 Feb 2017.
Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2009). *The Course of Study for Upper Secondary Schools*. Tokyo: MEXT.
Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2011). *The Course of Study for Elementary Schools*. Tokyo: MEXT.
Nakata, Y. (2010) Improving the clssroom language proficiency of non-native teachers of English: What and how? *RELC Journal, 41*(1), 76–90. doi:10.1177/0033688210362617
Nakata, Y., IKeno, O., Naganuma, N., Kimura, Y., & Andrews, S. (2012). Classroom ENglish language benchmarks for Japanese EFL teachers. *Proceedings of the JACET 51th international convension*, 20–27.
Newby, D. (2007). The European portfolio for student teacher of languages [Adobe Digital Editions version]. *Babylonia, 3*, 23–26. Retrieved from http://babylonia.ch/fileadmin/user_upload/documents/2007-3/newby.pdf.
North, B. (2009). *A profiling grid for language teachers.* Paper presented at the International Meeting on Training, Quality and Certification in Foreign Language Teaching, Siena, Italy. Retrieved from http://clients.squareeye.net/uploads/eaquals/North-%20TQAC.pdf
Orr, M. (2002). The FCE speaking tests: using rater reports to help interpret test scores. *System, 30*(2), 143–154. http://dx.doi.org/10.1016/S0346-251X(02)00002-7.
Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*(2), 219–233. doi:10.2307/3588333.
Pasternak, M., & Bailey, K. M. (2004). Preparing nonnative and native English-speaking teachers: Issues of professionalism and proficiency. In L. D. Kamhi-Stein (Ed.), *Learning and teaching from experience: Perspectives on nonnative English-speaking professionals* (pp. 155–175). Ann Arbor: The University of Michigan Press.
Rossner, R. (2009). *Methods of teacher assessment and the EAQUALS profiling grid for language teachers*. Paper presented at the International Meeting on Training, Quality and Certification in Foreign Language Teaching, Siena, Italy. Retrieved from http://clients.squareeye.net/uploads/eaquals/Rossner%20-%20Assessment%20and%20the%20EAQUALS%20Profiling%20Grid%20TQAC%202009.pdf
Sato, T. (2011). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing, 29*(2), 223–241. doi:10.1177/0265532211421162.
Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27–33. doi:10.1111/j.1540-4781.1992.tb02574.x.
Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.
Throndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25–29. doi:10.1037/h0071663.
Weigle, S. C. (1994). Effects of training on raters of ESL composition. *Language Testing, 11*(2), 197–223. https://doi.org/10.1177/026553229401100206.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205.
Wertsch, J. V. (1991). *Voices of the mind*. Cambridge: Harvard University Press.
Yorozuya, R., & Oller, J. W. (1980). Oral proficiency scales: construct validity and the halo effect. *Language Learning, 30*(1), 135–153.